Somatic and germline mutation periodicity follow the orientation of the DNA minor groove around nucleosomes

Oriol Pich[1], Ferran Muiños[1], Radhakrishnan Sabarinathan[1,&], Iker Reyes-Salazar[1], Abel Gonzalez-Perez[1,2,*], Nuria Lopez-Bigas[1,2,3,*,†]

Affiliations:

1. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain.

2. Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.

3. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

&. Present address: National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore 560065, India.

* Co-senior authors

†Corresponding author. E-mail: nuria.lopez@irbbarcelona.org

Summary

Mutation rates along the genome are highly variable and influenced by several chromatin features. Here we addressed how nucleosomes, the most pervasive chromatin structure, affect the generation of mutations. We discovered that within nucleosomes the somatic mutation rate across several tumor cohorts exhibit a strong 10 base-pair (bp) periodicity. This periodic pattern tracks the alternation of the DNA minor groove facing toward and away from the histones. The strength and phase of the mutation rate periodicity are determined by the mutational processes active in tumors. We uncovered similar periodic patterns in the genetic variation among human and *Arabidopsis* populations, also detectable in their divergence from close species, indicating that the same principles underlie germline and somatic mutation rates. We propose that differential DNA damage and repair processes dependent on the minor groove orientation in nucleosome-bound DNA significantly contribute to the 10 bp periodicity in AT/CG content in eukaryotic genomes.

Keywords

Tumor mutations, somatic mutations, germline variability, DNA damage, DNA repair, nucleosome positioning, WW periodicity

Introduction

Covering between 75% and 90% of the genomes of eukaryotes and several archaea, nucleosomes provide the first level of DNA compaction (Segal et al., 2006). Furthermore, due to the array of post-translational modifications that may be attached to histones, nucleosomes play an important role in the regulation of gene expression (McGinty and Tan, 2015).

DNA sequences wrap around nucleosomes with varying affinity (Thåström et al., 1999); those that more stably bound nucleosomes contain A/T di-nucleotides separated by 10 base pairs, or one DNA helix turn (Anselmi et al., 2000; McGinty and Tan, 2015; Satchwell et al., 1986). This is due to structural constrains, which favor the presence of A/T base pairs at stretches of the DNA with the minor groove facing the histones. Computational studies have suggested that the succession of A/T (or WW) di-nucleotides following a 10-bp periodicity (often termed WW periodicity) is one of the major determinants of the preservation of the *rotational* positioning –or orientation of the DNA relative to the histones core– of nucleosomes. On the other hand, the *translational* positioning of nucleosomes, that is, their location along the DNA fiber, is influenced by other sequence features, such as long homopolymeric sequences, located mainly at the start of linkers (Iyer, 2012; Kaplan et al., 2009; Langley et al., 2014; Liu et al., 2011; Struhl and Segal, 2013; Valouev et al., 2011). The WW periodicity related to the rotational positioning occurs along the genomes of organisms with nucleosomes, ranging from archaebacteria to higher eukaryotes (Herzel et al., 1998; Mrázek, 2010; Tolstorukov et al., 2011).

Recent studies of human somatic and germline variants have found large variability in the mutation rate along the genome. This variability correlates with certain genomic features, such as replication time (Stamatoyannopoulos et al., 2009), chromatin compaction (Schuster-Böckler and Lehner, 2012), expression level (Lawrence et al., 2013), the binding of transcription factors, the presence of nucleosomes, CTCF binding, and histone marks that distinguish exons from introns, among others (Chen et al., 2012; Frigola et al., 2017; Kaiser et al., 2016; Katainen et al., 2015; Morganella et al., 2016; Perera et al., 2016; Prendergast and Semple, 2011; Sabarinathan et al., 2016; Yazdi et al., 2015).

Here, to study the influence of the translational and rotational positioning of nucleosomes on the generation of the mutations before selection, we analyzed the somatic mutations observed across tumors, which are exposed to little selection once the few mutations driving tumorigenesis are filtered out (Frigola et al., 2017; Martincorena et al., 2017). We observed strong periodic patterns in the mutation rate of several cancer types, which track the alternance of nucleosomes and linkers and the rotational orientation of the minor groove of the DNA with respect to histones. We linked the orientation and magnitude of these mutation rate periodicities to the mutational processes present in tumor samples. Moreover, we demonstrated that the observed periodicities of the mutation rate are the result of a complex interaction between the processes of DNA damage and repair within the nucleosome territory. We then showed that spontaneous variation in human and *A. thaliana* populations also exhibit periodicity within nucleosome-occupied regions. Finally, we present a model that shows that the mutation rate

periodicity could have contributed to the development and maintenance of the WW periodicity in eukaryotic genomes across evolution.

Results

Periodic mutation rate tracks the alternation of nucleosome-covered and linker DNA

The presence of nucleosomes leads to a periodic pattern formed by alternating nucleosome-covered (147 bp; orange in model in Fig, 1a and subsequent figures) and linker (variable sizes; purple) DNA sequences (Voong et al., 2017). Previous studies by our group and others (Morganella et al., 2016; Sabarinathan et al., 2016) have shown that nucleosome-covered and linker DNA exhibit different somatic mutation rates in some tumor types. To study more systematically the influence of this periodic mutation rate, we first obtained the positions of nucleosomes along the genome of human lymphoblastoid cell lines mapped using the MNase cut efficiency (Gaffney et al., 2012). In each genomic region, defined by a peak of nucleosome density, we selected the dyad with more support from the MNase-seq experiment –assumed to correspond to the position with highest occupancy across cells. We retrieved whole-genome somatic mutations identified (Fredriksson et al., 2014; ICGC, 2010) across 3494 tumors from 28 cohorts (Methods; Table S1). We retained for further analyses those overlapping intergenic nucleosome-covered (henceforth, nucleosomes) and linker DNA, in order to minimize the potential effect of selection (see STAR Methods).

The rate of mutations observed in a 2001 bp-wide window centered at each nucleosome dyad across the samples of several cohort of tumors (Fig. 1 and Fig. S1a) exhibits a periodicity that tracks the alternation of nucleosome-covered and linker DNA, in agreement with previous findings (see above). We computed the difference between the observed mutation rate at each nucleotide across the stacked 2001 bp-wide windows (red wave-like signal in Fig. 1a) and its expected mutation rate (black signal), derived from the rate of changes across the genome in a penta-nucleotide context (Figs. 1a and S1b; STAR Methods). This difference is represented as a relative increase (i.e., taking the expected rate as baseline) of the observed mutation rate (two-color signal in Fig. 1a and Fig. 1b left bottom panel).

To characterize the periodic structure of the relative increase of the mutation rate, we constructed its periodogram (Fig 1b, right top panel), and computed the signal-to-noise ratio (SNR) of its maximum power period (MP). To assess the significance of the SNR, we computed the (expected) SNR of the relative increase of the mutation rate of each of 1000 permutations-based expected mutation rates. By counting the number of permutations with expected SNR above the observed SNR, we obtained an empirical p-value. These p-values were corrected to account for false discovery rate, thus yielding q-values (see STAR Methods).

The relative increase of the mutation rate computed for esophageal adenocarcinomas (Fig. 1b, right panel) exhibits its MP at 191.41 bp –the approximate length of a nucleosome-linker stretch of DNA– with a significant SNR of 300.45 (q-value<0.002; Table S2). To assess the orientation of the relative increase of the mutation rate signal, i.e., whether its maxima were located at the

nucleosomes or the linkers, we computed the phase shift of the relative increase of the mutation rate signal at period 191 bp with respect to a reference sinusoidal signal with maxima at the nucleosome dyads. Signals with a phase shift closer to 0 are deemed to have a phase of 1, whereas those exhibiting a phase shift closer to π are considered to have a phase of -1. For example, esophageal adenocarcinomas exhibit a relative increase of the mutation rate with phase 1 (Fig. 1b), while that of lung adenocarcinomas presents a phase of -1 (Fig. 1c,d). Cohorts with non-significant SNR (such as ovarian cancer in Fig. 1d, right top panel) do not exhibit any dominant period of the relative increase of the mutation rate.

The mutation rate in most cohorts with a significant SNR exhibit a phase 1 (top panel in Fig. 1c), that is, with maxima –the greatest excess of mutation rate over its expected value– located within nucleosomes (as the melanomas in Fig. 1d left top panel). Esophageal adenocarcinomas, melanomas and gastric adenocarcinomas show the most significant relative increase of mutation rate (Fig. 1c). Only lung adenocarcinomas and squamous cell carcinomas and uterine adenocarcinomas show a relative increase of mutation rate that peaks at the linkers (Fig. 1c, bottom panel; Fig. 1d, left bottom panel).

Periodic mutation rate tracks minor groove orientation within the nucleosomes

The DNA wrapping a nucleosome also exhibits an alternating pattern of structurally distinct stretches: ~10-bp interspersed segments of DNA with the minor groove facing the histones (green in model in Fig, 2a and subsequent figures) and away from them (yellow). Certain rotational positions are preferred even by nucleosomes with lowly conserved translational positioning across cells (Iyer, 2012; Kaplan et al., 2009; Langley et al., 2014; Liu et al., 2011). We hypothesized that the structural differences of these stretches of DNA could influence the mutation rate.

We computed the observed mutation rate at each nucleotide across a 117-bp window centered at the dyad of each nucleosome, the length of a nucleosome core after removal of nucleosome-linker boundaries (Gaffney et al., 2012). We subtracted the expected mutation rate at each nucleotide within the stacked 117-bp sequences (as in the previous section), thus obtaining a relative increase of the mutation rate. As with the nucleosome-linker pattern, we then determined the periodicity of the relative increase of mutation rate via the periodogram, and its magnitude through the SNR of the MP. We computed an empirical p-value of the observed SNR by comparing it to the SNR computed for permutation-based expected mutation rates. Finally, to determine the phase of the relative increase of mutation rate, we computed the phase shift of the signal at period 10.3 bp with respect to a reference sinusoidal signal with maxima at inwardly-facing minor grooves. If the relative increase of mutation rate yielded a phase shift close to 0 it was assigned a phase of 1 (esophageal adenocarcinomas in Fig. 2b, with SNR=613.64 and q-value<0.003), whereas those with a phase shift closer to π exhibited a phase of -1 (see STAR Methods).

The mutation rate of the tumors of several cohorts showed a strong periodic pattern following the orientation of the minor groove with respect to histones (Fig. 2b, c, d, S2a and b; Table S2). The relative increase of mutation rate of all cohorts with significant SNR showed an MP around

10 bp, or one helix turn (Fig. 2c). In some tumors (e.g., esophageal adenocarcinomas in Fig. 2b, and malignant lymphomas in Fig. 2d, right bottom panel) the relative increase of mutation rate peaked at stretches of DNA with the minor groove facing the histones (phase 1). In others (e.g., melanomas and lung adenocarcinomas), its maxima were at stretches of DNA with the minor groove facing away from the histones (phase -1).

If this periodicity of mutation rate was originated by different structural properties of DNA stretches, we expected that nucleosomes with stronger rotational positioning showed more strongly periodic relative increase of mutation rate. To test this, we separated the group of nucleosomes with the highest score of rotational setting (strong rotational position), and a group with equal number of nucleosomes at the lower end of scores of rotational setting (weak rotational position). Across most cohorts, strong rotationally positioned nucleosomes exhibited greater SNR of the periodicity of the relative increase of mutation rate than weakly positioned ones (Fig. S2c). The trend for mutations to accumulate more than expected at stretches of DNA with the minor groove facing the histones or away from them is directly proportional to the strength of the rotational positioning of nucleosomes (see STAR Methods).

In summary, we observed strong periodicity in the mutation rate tracking the alternation of DNA minor groove facing toward the histones and away from them, the orientation of which varies between tumor types.

Tumors with different mutational signatures show dissimilar mutation rate periodicity

We hypothesized that various mutational processes active across cohorts could be responsible for the differences in the magnitude and orientation of the periodicity of the relative increase of mutation rate. Since the mutations of different tumors may be exposed to distinct mutational processes, we computed the relative increase of mutation rate within nucleosome-covered DNA for each individual tumor with at least 500 mutations overlapping nucleosomes.

We reduced the signal of the relative increase of mutation rate to a single number (minor-in), computed from the sum of the values corresponding to the three nucleotides at the center of the DNA segments with the minor groove facing the histones. Tumors with positive minor-in relative increase of mutation rate possess higher-than-expected mutation rate at these stretches of DNA. We then computed an empirical p-value to assess the significance of the SNR of the relative increase of mutation rate of each tumor, as explained in the previous section (STAR Methods).

In Figure 3a, cohorts appear sorted in ascending order of the median minor-in relative increase of mutation rate across tumors. Although most melanomas (extreme left of the graph; median minor-in relative increase of mutation rate, -0.075) have a significantly negative value, one of its samples bear a significantly positive minor-in relative increase of mutation rate (0.1), comparable to that of most esophageal adenocarcinomas (extreme right; median minor-in relative increase of mutation rate, 0.085). To understand the differences between tumors, we deconstructed the contribution of mutational signatures (Alexandrov et al., 2013; Rosenthal et

al., 2016) to each of them (Fig. 3b). A representative melanoma (second-to-last pie-chart), as well as the cells of a normal skin sample (fourth pie-chart), with the majority of mutations following the UV signature 7 exhibit a significant negative minor-in relative increase of mutation rate, as do lung adenocarcinomas, with major contribution of signature 4 (second pie-chart). On the other hand, the melanoma sample with significant positive minor-in relative increase of mutation rate has a major contribution of signature 17 (first pie-chart). The shape and orientation of the relative increase of mutation rate of this melanoma is very similar to that of esophageal and stomach adenocarcinomas (third and sixth pie-charts). Tumors with contributions from signatures 14 and 10, such as POLE-mutant uterine and colorectal adenocarcinomas (third-to-last and last pie-charts) also exhibit a positive minor-in relative increase of mutation rate. In summary, specific mutational processes are the major determinants of the direction of the mutation rate periodicity within nucleosome covered DNA in individual tumors.

The mutational processes active in tumors influence the orientation of their mutation rate periodicity

We then sought to comprehensively delineate the association between mutational processes and the magnitude and orientation of the relative increase of mutation rate, including the potential role of signatures that are minor contributors to the mutations of each tumor, but are active across many samples. We deconstructed the contribution of each signature to the mutational landscape of the 505 tumors in the TCGA dataset (with a unified calling process), and pooled all mutations contributed by each signature (STAR Methods). (Two algorithms used for the signature deconstruction yielded highly concordant results; Fig. S3a). In the alternation of nucleosomes and linkers, mutations contributed by signatures 1, 7, 17 and 18 yield a relative increase of mutation rate with significant SNR with phase 1 (higher mutation rate than expected in nucleosomes; Fig. S3b and S3c). On the other hand, the relative increase of mutation rate of those contributed by signatures 4, 6 and 16 have a significant SNR with phase -1.

With respect to the periodicity tracking the orientation of the minor groove of the DNA within the nucleosome, mutations contributed by signature 7 –resulting from DNA adducts formed by the action of UV light (Ikehata and Ono, 2011; Yu and Lee, 2017)– yield a strongly periodic relative increase of mutation rate, with significant SNR of 207.02 and a phase of -1 (Fig. 4a, b). The periodicity of mutations contributed by signatures 4 –associated to DNA adducts generated by tobacco carcinogens (Alexandrov et al., 2016)– follows the same orientation as signature 7 (phase=-1; SNR=78.32; Fig. 4a,b). Mutations contributed by signature 10, active in POLE mutant tumors (Haradhvala et al., 2018), and signature 17, linked to the oxidation of guanines in the nucleotide pool (Tomkova et al., 2017), show a greater accumulation than expected in stretches of DNA with minor groove facing the histones (phase=1) (Fig. 4a,b). The SNR of the periodicity of the relative increase of the rate of mutations contributed by several signatures (Fig. 4c) is greater for nucleosomes with strong rotational position than those with weak rotational position. Finally, the periodicity of the mutations contributed by several of these signatures was also detected across exonic nucleosomes (Fig. 4d), demonstrating that the frequency of somatic coding mutations is also affected by structural differences of the DNA within nucleosomes.

A complex interplay between DNA damage and repair within nucleosomes

We next asked which of the processes involved in mutational signatures –DNA damage or repair– is ultimately responsible for the observed periodicity of the mutation rate within nucleosomes (Figs. 5 and S4).

We studied, as a model, the generation and repair of UV-induced lesions, namely, cyclobutane pyrimidine dimers, or CPDs, and 6,4 pyrimidine-pyrimidones, or (6-4)PPs (Marteijn et al., 2014; Osakabe et al., 2015; Yu and Lee, 2017) within nucleosomes. To this end, we superimposed the whole-genome map of CPDs and (6-4)PPs detected in a cell line of human fibroblasts at different time points after treatment with UV light (Hu et al., 2017) to the coordinates of nucleosomes (blue signal in Fig. 5a and S4a). We then calculated the relative increase of the rate of each type of adduct (purple signal Fig. 5a) using the expected relative damage frequencies at Cs or Ts within each possible penta-nucleotide context computed from the whole-genome damage maps (as explained above for mutations). Finally, we obtained the magnitude and phase of its periodicity. In agreement with previous reports (Hara et al., 2000), we observed that the rate of both, CPDs (Fig. 5b, left panels) and (6-4)PPs (Fig. S4a) follows a strongly periodic pattern (MP=10.15 bp, SNR=112.69), with maxima at stretches of DNA with the minor groove facing away from histones (phase=-1). CPDs formed within nucleosomes with strong rotational position possess stronger periodicity than those with weak rotational position (right barplot). In contrast, the formation of both types of lesions on naked DNA follows the di-pyrimidine content, and is out-of-phase with respect to the periodicity observed in nucleosome-covered DNA (Fig. S4b).

We then addressed whether the nucleotide excision repair (NER) –charged with UV-damage correction (Schärer, 2013)– shows a differential activity within nucleosome-covered DNA. We estimated the rate of repair at each position by computing the proportion of damaged sites detected immediately after UV irradiation that remain after 24h. The repair of both, CPDs (Fig. 5b) and (6-4)PPs (Fig. S4c) is periodic, as observed in yeast (Mao et al., 2017). CPDs and (6-4)PPs at DNA stretches with the minor groove facing away from histones are repaired at higher rates than their counterparts at sites of minor grooves facing nucleosomes (CPDs repair SNR=29.5; phase=-1). The repair of CPDs shows greater periodicity within nucleosomes with strong rotational position than within those with weak rotational position (right barplot). Taken together, these results indicate that the periodicity of UV-induced mutation rate within nucleosomes (Fig. 4b, right bottom panel), is driven by the periodicity in the rate of accumulation of CPDs and (6-4)PPs, rather than by that of their repair. This is further supported by the observation that the mutation rate periodicity in cells lacking global NER (XPC mutants) is very similar to that of their XPC wild-type counterparts (Fig. S4d)

We carried out the same type of analysis to compute the periodicity of methyl-guanines (MeGs) generated in yeast DNA as a result of exposure to methyl methanesulfonate (MMS) (Mao et al., 2017). In contrast to UV damage, MMS-induced MeGs do not exhibit a significant periodic signal around 10 bp (Fig. 5c; SNR=16.11, MP=11.36; p-value=0.977). Upon analysis of the rate

of repair of these MeGs (2h after exposure), as explained above for UV damage, we found a strong periodicity (SNR=97.31) with peaks at sites of minor grooves facing away from histones (Fig. 5c, right panel), as previously reported (ref Mao). In other words, like NER, the base excision repair (BER) pathway –involved in the correction of MeGs (Krokan and Bjørås, 2013; Wallace, 2014)– exhibits higher efficiency at minor-out segments (Rodriguez and Smerdon, 2013).

To study whether the increased DNA repair efficiency of BER and NER at sites of minor grooves facing away from histones is related to improved accessibility, we computed the relative increase of DNAse nicks –a good proxy of minor groove accessibility– within nucleosome-covered DNA sequences. The relative increase of DNAse cutting efficiency (Fig. 5d) peaks at minor grooves facing away from histones in both human (SNR=97.47; phase=-1) and yeast nucleosomes (SNR=263.87; phase=-1), which had been previously reported (Mao et al., 2017).

Taken together these results indicate that the periodic structure and orientation of the relative increase of the mutation rate within nucleosome-covered DNA is determined by the combined effects of DNA damage and repair efficiency at DNA stretches with the minor groove facing toward and away from histones.

Periodic germline variation and interspecific divergence also track alternative orientations of the minor groove

We have shown that somatic mutations contributed by certain mutational processes are generated at uneven rates at segments of DNA with the minor groove facing inwards or outwards. Since DNA repair machineries (and some DNA damages) involved in these mutational processes are also relevant in germ cells, we speculated that the same type of periodicity described for somatic mutations could be observed for spontaneous germline variants and interspecies divergence.

We mapped 8.9 and 3.7 millions of rare variants (allelic frequency<0.01) observed across human (Gibbs et al., 2015) and *A. thaliana* (Alonso-Blanco et al., 2016) populations, respectively onto the intergenic nucleosomes of either species. We then computed the observed and expected (as explained for mutations) rates of these variants within nucleosome-covered DNA. The relative increase of the rate of rare variants amongst both human (SNR=59.55; p-value<0.001) and *Arabidopsis* (SNR=31.44; p-value<0.001) individuals is significantly periodic (MP=10.15), with the maxima at stretches of DNA with the minor groove facing histones (Figs. 6a and S5). The SNR of this periodicity is higher within human nucleosomes with strong rotational position (right barplot).

We then analyzed the distribution of diverging genomic sites for these species (i.e. humans and *A. thaliana)* within DNA wrapped around nucleosomes with strong rotational position. From the three-way whole-genome alignment of the species under study and two species close to each of them (*P. troglodytes* and *G. gorilla* in the case of *H. sapiens; A. lyrata and B. rapa* for *A. thaliana*), we identified divergent sites. On detail, we annotated genomic sites that have changed from a C to a T –representing a majority of *de novo* genomic variants across many

eukaryotic species– in the divergence between each species under analysis and the common ancestor of the trio, while remaining unchanged in the other two (i.e., polarized divergence) (Langley et al., 2014). Furthermore, we required that both nucleotides flanking each annotated C>T variant were identical across the three species (STAR Methods). We found a significantly periodic relative increase of the rate of polarized C>T divergence in both humans (SNR=39.24) and *A. thaliana* (SNR=54.6). In both species the stretches of DNA with the minor groove facing histones possess greater-than-expected rate of C>T divergence.

These results indicate that germline variation within nucleosomes occurs at different rates at DNA with minor groove facing toward or away from the histones. The higher rate of spontaneous variants and interspecies C>T divergent sites observed at DNA stretches with the minor groove facing the nucleosome, coincides with that of somatic mutations contributed by several mutational processes.

A role for differential mutation rate in genomic sequence periodicity

Since nucleosomes cover between 75% and 90% of eukaryotic DNA, it is reasonable to speculate that the 10-bp periodicity in the rate of inherited variants within nucleosome-covered DNA could affect the composition of the genome across evolutionary time. Actually, the alternation of stretches of DNA with the minor groove facing the histones and away from them are enriched for AT and GC di-nucleotides, shaping the 10-bp periodicity of A/Ts, or WW periodicity (Iyer, 2012; Kaplan et al., 2009; Liu et al., 2011). The WW periodicity becomes apparent when the nucleosome-covered DNA sequences –centered at their dyads– of five eukaryotic genomes are stacked (Brogaard et al., 2012; Gaffney et al., 2012; Langley et al., 2014; Voong et al., 2016; Zhang et al., 2015), (Fig. 7a; different color legend than previous figures). It is reminiscent of that observed for the mutations contributed by certain mutational processes, such as those represented by signatures 17, 10, 14, and 18, and also to that of germline variants and interspecies C>T divergence, i.e, maxima at minor-in segments (phase 1).

We designed a general approach based on methods developed by others to assess the strength of the periodicity of WW dinucleotides across the genomes of organisms without experimentally mapped nucleosomes (Herzel et al., 1999; Mrázek, 2010) (Fig. 7b). We first divided the genomes of 76 eukaryotes (Table S3) into 1 Mb chunks. In each chunk, we tracked the positional composition of di-nucleotide pairs and computed an autocorrelation score as the ratio between the frequency of observed pairs of WW dinucleotides at a given distance and their frequency inferred from the dinucleotide composition of the chunk (left panel). Any consistent periodicity of WW dinucleotides in the DNA as a categorical signal (henceforth, walk periodicity) would appear as a periodic component of the same period in the autocorrelation curve (blue signal), which we smoothed (green signal) to filter out the genomic widespread 3-bp periodicity. We did the same analysis for 100 randomly generated DNA sequences of the same length produced using a Markov Chain Monte Carlo sampling, maintaining the underlying di-nucleotide composition and adjacency frequencies of the real chunk. Next, we computed the fold-power increase (FPI) –and an associated p-value– of each period detected in the autocorrelation function (second panel) with respect to the behavior of the random sequences, and obtained the

distributions of FPI for each period across chunks (third panel). Finally, to determine if a particular genome shows any detectable periodicity of WW dinucleotides, we counted the number of its chunks with significant FPI at each period between 6 bp and 19 bp, and computed the bias towards the 10 bp period (odds-ratio) with respect to uniformity (last panel; STAR Methods).

We thus obtained the 10-bp odds-ratio and associated p-value of the periodic structure of WW dinucleotides across chunks for each of the 76 eukaryotic genomes (Fig. 7c; Table S4). Most genomes analyzed show an enrichment for 10-bp WW dinucleotides (odds-ratio>1) with a wide range of variability among them; all other periods between 6 bp and 19 bp show much lower enrichment across most genomes (Fig. S6; Table S4). While in some organisms this enrichment is 100-fold and the vast majority of their genomic chunks exhibit a clear MP at 10 bp (e.g. *S. cerevisiae*), in others this fraction is less than 10% and the enrichment is more subtle (e.g. *H. sapiens, M. musculus*, and *D. melanogaster*), even though a strong WW periodicity is apparent across stacked nucleosomes (Jin et al., 2016) (Fig. 7a).

Having confirmed the widespread WW periodicity across eukaryotic genomes, it is necessary to explain how this periodic pattern arose. We hypothesized that the observed periodicity of *de novo* variants, and inter-species divergence, resulting from the interaction of mutational processes with the nucleosome structure, could have a role in the development and maintenance of the WW periodicity (Fig. 7d). The generation of an excess of *de novo* biased variants –i.e., more variants yielding A/Ts– at sites of minor grooves facing histones, would leave these sites enriched for these nucleotides over evolutionary time. *De novo* mutations detected in many species exhibit this bias due to the spontaneous deamination of 5-methylcytosine (Long et al., 2018; Cooper et al., 2010; Pfeifer, 2017; Shen et al., 1994).

To illustrate that the WW periodicity could have arisen given these conditions, we devised a mutational simulator (Fig. 7e). It begins with a completely random 1 million bp DNA sequence with nucleosomes. It receives *de novo* mutations at a fixed rate with a minor-in relative increase of the mutation rate of 4% (similar to that computed for several tumors; Fig. 3a) and the mutational spectra of *de novo* variants across humans (including a C>T bias of 80%) (Long et al., 2018). After each mutational iteration, the SNR of the WW periodicity of the sequence is computed. The SNR of the WW periodicity begins to grow after roughly 3000 iterations of the simulator (yellow line). If either of the two conditions (periodicity and biased variants) is not met (blue and purple broken lines) the WW periodicity does not appear within 6000 iterations (see STAR Methods).

In summary, we confirm the widespread WW periodicity across eukaryotic genomes and demonstrate that in the absence of other evolutionary forces, the periodic *de novo* mutation mutation rate and the C>T mutation bias would be sufficient for its development.

Discussion

Understanding how mutations are generated along the genome is key to explaining how evolution has shaped its sequence. Here we demonstrate that the positioning of nucleosomes, the most pervasive chromatin feature, affects the mutation rate in both somatic and germ cells. We found that the differences in structural features of nucleosome-covered DNA and linkers, and between stretches of DNA with the minor groove facing histones and away from them –once their dissimilar sequence compositions are accounted for– result in unequal rates of mutation generation in cells.

We have employed the positions of nucleosomes obtained by a high-quality available mapping of human nucleosomes in lymphoblastoid cells (Gaffney et al., 2012) to map somatic mutations and germline variants obtained from different cell types. We expect that the position of intergenic nucleosomes is less variable than that of nucleosomes located around active promoters, related to cell identity. The observed periodicity of mutation rate across different tumor types may be regarded as a validation of the usefulness of mapped nucleosomes across cell types. We envision that future work with mutations and nucleosomes mapped in the same cell type will produce even clearer periodic patterns. While we focus the discussion on the tumor types and signatures with most salient periodic patterns, others that are now detected as borderline significant will probably show –with more accurate nucleosome maps and larger cohorts sequenced (see results of separate cohorts in Fig. S3c)– clearer periodic patterns.

Mutations arise from unrepaired DNA lesions, produced by mutagenic agents, or mismatches, which result in the incorporation of incorrect nucleotides by DNA polymerases during replication (Alexandrov et al., 2013; Francioli et al., 2015; Haradhvala et al., 2018; Hollstein et al., 2016; Stephens et al., 2012; Tomkova et al., 2017). Our results indicate that the interaction between different mutagenic agents and DNA repair mechanisms with the nucleosome core determines the appearance of mutation rate periodicities within nucleosomes. For example UV-induced mutations (signature 7) appear at higher-than-expected rates at stretches of DNA with the minor groove facing away from nucleosomes because CPDs and (6-4)PPs are formed at higher rates at these sites, rather than the differences in global NER activity. On the other hand, the periodicity of signature 17 mutations –probably a result of oxidative damage (Tomkova et al., 2017), repaired at least in part by the BER– could appear due to decreased repair of oxidized guanines at minor-in sites.

Although the pervasive WW periodicity of eukaryotic genomes (Iyer, 2012; Kaplan et al., 2009; Langley et al., 2014; Liu et al., 2011) (Fig. 7a, c) has been linked to the favorable DNA bending around nucleosomes, its emergence and maintenance are not completely understood. On the basis of our results, we hypothesize that DNA damage and the mechanisms of DNA repair active in germ cells, in their interaction with DNA wrapped around nucleosomes give rise to a 10-bp periodicity in the rate of *de novo* variants and interspecies divergence. Thus, the appearance of nucleosomes during evolution and their positioning along the genomic sequence may have favored the development of a higher-than-expected occurrence of C>T *de novo* variants –due to underlying mutational process– at stretches of DNA with minor groove facing the histones. We propose that the generation and maintenance of the WW periodicity may be explained, at least in part, by the periodicity of *de novo* biased genetic variants thus arisen.

Author contributions

O.P. analyzed the periodicity of somatic and germline mutation rates, produced most of the figures, participated in the design of analyses and in the interpretation of results, and edited the manuscript. F.M. developed the statistical framework to analyze periodic patterns studied, produced figure 7b and c, participated in the design of the analyses and in the interpretation of results, and edited the manuscript. O.P, F.M. and I.R.-S. developed and carefully tested the reproducibility of computational code employed in the analyses. S.R. participated in the analyses and in the interpretation of results. A.G.-P. participated in the design of analyses and in the interpretation of results, oversaw the study and drafted and edited the manuscript. N.L.-B. conceived the study, participated in the design of analyses and in the interpretation of results, oversaw the study and drafted and edited the manuscript.

Acknowledgments

N.L-B. acknowledges funding from the European Research Council (consolidator grant 682398) and Spanish Ministry of Economy and Competitiveness (SAF2015-66084-R, MINECO/FEDER, UE). IRB Barcelona is a recipient of a Severo Ochoa Centre of Excellence Award from the Spanish Ministry of Economy and Competitiveness (MINECO; Government of Spain) and is supported by CERCA (Generalitat de Catalunya). O.P. is supported by the BIST and FI PhD Fellowship. A.G-P. is supported by a Ramón y Cajal contract (RYC-2013-14554). The results shown here are in whole or part based upon data generated by the TCGA Research Network.

Declaration of Interests

The authors declare no competing interests.

References

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A. V, Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alexandrov, L.B., Ju, Y.S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., et al. (2016). Mutational signatures associated with tobacco smoking in human cancer. Science *354*, 618–622.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K.M., Cao, J., Chae, E., Dezwaan, T.M., Ding, W., et al. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell *166*, 481–491.

Anselmi, C., Bocchinfuso, G., De Santis, P., Savino, M., and Scipioni, A. (2000). A Theoretical Model for the Prediction of Sequence-Dependent Nucleosome Thermodynamic Stability. Biophys. J. *79*, 601–613.

Brogaard, K., Xi, L., Wang, J.-P., and Widom, J. (2012). A map of nucleosome positions in yeast at base-pair resolution. Nature *486*, 496–501.

Chen, X., Chen, Z., Chen, H., Su, Z., Yang, J., Lin, F., Shi, S., and He, X. (2012). Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. Science (80-.). *335*, 1235–1238.

Cooper, D.N., Mort, M., Stenson, P.D., Ball, E. V, and Chuzhanova, N.A. (2010). Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. Hum. Genomics *4*, 406–410.

Cui, F., and Zhurkin, V.B. (2010). Structure-based analysis of DNA sequence patterns guiding nucleosome positioning in vitro. J. Biomol. Struct. Dyn. *27*, 821–841.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390–394.

Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., McLellan, M., et al. (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst. *6*, 271–281.e7.

Fickett, J.W., and Tung, C.S. (1992). Assessment of protein coding measures. Nucleic Acids Res. *20*, 6441–6450.

Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., van Duijn, C.M., Swertz, M., Wijmenga, C., van Ommen, G., et al. (2015). Genome-wide patterns and properties of de novo mutations in humans. Nat. Genet. *47*, 822–826.

Fredriksson, N.J., Ny, L., Nilsson, J.A., and Larsson, E. (2014). Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nat. Genet. *46*, 1258–1263.

Frigola, J., Sabarinathan, R., Mularoni, L., Muiños, F., Gonzalez-Perez, A., and López-Bigas, N. (2017). Reduced mutation rate in exons due to differential mismatch repair. Nat. Genet. *49*, 1684–1692.

Fujita, P. a, Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., et al. (2011). The UCSC Genome Browser database: update 2011. Nucleic Acids Res. *39*, D876-82.

Gaffney, D.J., McVicker, G., Pai, A.A., Fondufe-Mittendorf, Y.N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y., and Pritchard, J.K. (2012). Controls of Nucleosome Positioning in the Human Genome. PLoS Genet. *8*, e1003036.

Gibbs, R.A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J.G., Zhu, Y., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Gori, K., and Baez-Ortega, A. (2018). sigfit: flexible Bayesian inference of mutational signatures. BioRxiv 372896.

Hara, R., Mo, J., and Sancar, A. (2000). DNA damage in the nucleosome core is refractory to repair by human excision nuclease. Mol. Cell. Biol. *20*, 9173–9181.

Haradhvala, N.J., Kim, J., Maruvka, Y.E., Polak, P., Rosebrock, D., Livitz, D., Hess, J.M., Leshchiner, I., Kamburov, A., Mouw, K.W., et al. (2018). Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. Nat. Commun. *9*, 1746.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

Herzel, H., Weiss, O., and Trifonov, E.N. (1998). Sequence Periodicity in Complete Genomes of Archaea Suggests Positive Supercoiling. J. Biomol. Struct. Dyn. *16*, 341–345.

Herzel, H., Weiss, O., and Trifonov, E.N. (1999). 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics *15*, 187–193.

Hollstein, M., Alexandrov, L., Wild, C., Ardin, M., and Zavadil, J. (2016). Base changes in tumour DNA have the power to reveal the causes and evolution of cancer. Oncogene *36*, 158–167.

Hu, J., Adebali, O., Adar, S., and Sancar, A. (2017). Dynamic maps of UV damage formation and repair for the human genome. Proc. Natl. Acad. Sci. *114*, 201706522.

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95.

ICGC (2010). International network of cancer genome projects. Nature *464*, 993–998.

Ikehata, H., and Ono, T. (2011). The mechanisms of UV mutagenesis. J. Radiat. Res. *52*, 115–125.

Iyer, V.R. (2012). Nucleosome positioning: bringing order to the eukaryotic genome. Trends Cell Biol. *22*, 250–256.

Jin, H., Rube, H.T., and Song, J.S. (2016). Categorical spectral analysis of periodicity in nucleosomal DNA. Nucleic Acids Res. *44*, 2047–2057.

Kaiser, V.B., Taylor, M.S., and Semple, C.A. (2016). Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. PLoS Genet. *12*, e1006207.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., et al. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362–366.

Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A.E., Ristolainen, H., Hänninen, U.A., Cajuso, T., et al. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. Nat. Genet. *47*, 818–821.

Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics *26*, 2204–2207.

Krokan, H.E., and Bjørås, M. (2013). Base excision repair. Cold Spring Harb. Perspect. Biol. *5*, a012583.

Langley, S.A., Karpen, G.H., and Langley, C.H. (2014). Nucleosomes Shape DNA Polymorphism and Divergence. PLoS Genet. *10*, 25–27.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G. V, Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liu, H., Lin, S., Cai, Z., and Sun, X. (2011). Role of 10–11bp periodicities of eukaryotic DNA sequence in nucleosome positioning. Biosystems *105*, 295–299.

Long, H., Sung, W., Kucukyildirim, S., Williams, E., Miller, S.F., Guo, W., Patterson, C., Gregory, C., Strauss, C., Stone, C., et al. (2018). Evolutionary determinants of genome-wide nucleotide composition. Nat. Ecol. Evol. *2*, 237–240.

Mao, P., Brown, A.J., Malc, E.P., Mieczkowski, P.A., Smerdon, M.J., Roberts, S.A., and Wyrick, J.J. (2017). Genome-wide maps of alkylation damage, repair, and mutagenesis in yeast reveal mechanisms of mutational heterogeneity. Genome Res. *27*, 1674–1684.

Marteijn, J.A., Lans, H., Vermeulen, W., and Hoeijmakers, J.H.J. (2014). Understanding nucleotide excision repair and its roles in cancer and ageing. Nat. Rev. Mol. Cell Biol. *15*, 465–481.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D.C., Fullam, A., Alexandrov, L.B., Tubio, J.M., et al. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. Science (80-.). *348*, 880–886.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell *171*, 1029–1041.e21.

McGinty, R.K., and Tan, S. (2015). Nucleosome structure and function. Chem. Rev. *115*, 2255–2273.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. 51–56.

Morganella, S., Alexandrov, L.B., Glodzik, D., Zou, X., Davies, H., Staaf, J., Sieuwerts, A.M., Brinkman, A.B., Martin, S., Ramakrishna, M., et al. (2016). The topography of mutational processes in breast cancer genomes. Nat. Commun. *7*, 11383.

Mrázek, J. (2010). Comparative analysis of sequence periodicity among prokaryotic genomes points to differences in nucleoid structure and a relationship to gene expression. J. Bacteriol. *192*, 3763–3772.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics *28*, 1919–1920.

Oliphant, T.E. (2006). Guide to NumPy (Open Source Book).

Osakabe, A., Tachiwana, H., Kagawa, W., Horikoshi, N., Matsumoto, S., Hasegawa, M., Matsumoto, N., Toga, T., Yamamoto, J., Hanaoka, F., et al. (2015). Structural basis of pyrimidine-pyrimidone (6-4) photoproduct recognition by UV-DDB in the nucleosome. Sci. Rep. *5*, 1–5.

Perera, D., Poulos, R.C., Shah, A., Beck, D., Pimanda, J.E., and Wong, J.W.H. (2016). Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. Nature *532*, 259–263.

Perez, F., and Granger, B.E. (2007). IPython: A System for Interactive Scientific Computing. Comput. Sci. Eng. *9*, 21–29.

Pfeifer, G.P. (2017). DNA Methylation and Mutation. In ELS, (Chichester, UK: John Wiley & Sons, Ltd), pp. 1–5.

Pohl, A., and Beato, M. (2014). bwtool: a tool for bigWig files. Bioinformatics *30*, 1618–1619.

Prendergast, J.G.D., and Semple, C.A.M. (2011). Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. Genome Res. *21*, 1777–1787.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rodriguez, Y., and Smerdon, M.J. (2013). The structural location of DNA lesions in nucleosome core particles determines accessibility by base excision repair enzymes. J. Biol. Chem. *288*, 13863–13875.

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S., and Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. *17*, 31.

Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., and López-Bigas, N. (2016). Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature *532*, 264–267.

Satchwell, S.C., Drew, H.R., and Travers, A.A. (1986). Sequence periodicities in chicken nucleosome core DNA. J. Mol. Biol. *191*, 659–675.

Schärer, O.D. (2013). Nucleotide excision repair in eukaryotes. Cold Spring Harb. Perspect. Biol. *5*, a012609.

Schuster-Böckler, B., and Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature *488*, 504–507.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.P.Z., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778.

Shen, J.-C., Rideout, W.M., and Jones, P.A. (1994). The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res. *22*, 972–976.

Sokal R. (1995). Biometry the principles and practice of statistics in biological research Robert R. Sokal and F. James Rohlf. [electronic resource] - Version details - Trove.

Stamatoyannopoulos, J.A., Adzhubei, I., Thurman, R.E., Kryukov, G. V, Mirkin, S.M., and Sunyaev, S.R. (2009). Human mutation rate associated with DNA replication timing. Nat. Genet. *41*, 393–395.

Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R., et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature *486*, 400–404.

Stoica, P., and Moses, R. (2004). SPECTRAL ANALYSIS OF SIGNALS (Prentice Hall, New Jersey).

Struhl, K., and Segal, E. (2013). Determinants of nucleosome positioning. Nat. Struct. Mol. Biol. *20*, 267–273.

Thåström, A., Lowary, P.., Widlund, H.., Cao, H., Kubista, M., and Widom, J. (1999). Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. J. Mol. Biol. *288*, 213–229.

Tolstorukov, M.Y., Volfovsky, N., Stephens, R.M., and Park, P.J. (2011). Impact of chromatin structure on sequence variability in the human genome. Nat. Struct. Mol. Biol. *18*, 510–515.

Tomkova, M., Tomek, J., Kriaucionis, S., and Schuster-Boeckler, B. (2017). Widespread impact of DNA replication on mutational mechanisms in cancer. BioRxiv 111302.

Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z., and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. Nature *474*, 516–522.

Voong, L.N., Xi, L., Sebeson, A.C., Xiong, B., Wang, J.-P., and Wang, X. (2016). Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping. Cell *167*, 1555–1570.e15.

Voong, L.N., Xi, L., Wang, J.P., and Wang, X. (2017). Genome-wide Mapping of the Nucleosome Landscape by Micrococcal Nuclease and Chemical Mapping. Trends Genet. *33*, 495–507.

Wallace, S.S. (2014). Base excision repair: a critical player in many games. DNA Repair (Amst). *19*, 14–26.

Yazdi, P.G., Pedersen, B.A., Taylor, J.F., Khattab, O.S., Chen, Y.H., Chen, Y., Jacobsen, S.E., and Wang, P.H. (2015). Increasing nucleosome occupancy is correlated with an increasing mutation rate so long as DNA repair machinery is intact. PLoS One *10*, 1–16.

Yu, S.-L., and Lee, S.-K. (2017). Ultraviolet radiation: DNA damage, repair, and human disorders. Mol. Cell. Toxicol. *13*, 21–28.

Zentner, G.E., and Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. Nat. Struct. Mol. Biol. *20*, 259–266.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Res. *46*, D754–D761.

Zhang, T., Zhang, W., and Jiang, J. (2015). Genome-Wide Nucleosome Occupancy and Positioning and Their Impact on Gene Expression and Evolution in Plants. Plant Physiol. *168*, 1406–1416.

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., and Wang, L. (2014). CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics *30*, 1006–1007.

Zheng, C.L., Wang, N.J., Chung, J., Moslehi, H., Sanborn, J.Z., Hur, J.S., Collisson, E.A., Vemula, S.S., Naujokas, A., Chiotti, K.E., et al. (2014). Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes. Cell Rep. *9*, 1228–1234.

Zhong, J., Luo, K., Winter, P.S., Crawford, G.E., Iversen, E.S., and Hartemink, A.J. (2016). Mapping nucleosome positions using DNase-seq. Genome Res. *26*, 351–364.

Figure legends

Figure 1. Periodicity of tumor mutation rate across nucleosomes and linkers

(a) Alternation of nucleosomes and linkers in DNA. The nucleotide-wise observed and expected mutation rate in 2001 bp sequences are computed (red and black signals, respectively).

(b) Observed and expected mutation rate of esophageal adenocarcinomas (top-left); relative increase of mutation rate (bottom-left) and its periodogram (right). Vertical broken lines represent dyads.

(c) Signal-to-noise ratio (SNR; y-axis) of the strongest period (x-axis) in the relative increase of mutation rate of cohorts with phase 1 (top panel) or -1 (bottom panel). Cohorts with a relative increase of mutation rate with SNR>8 and q-value<0.05 are circled.

(d) Four examples of periodicity of the relative increase of the mutation rate. Clockwise from top left corner: skin melanomas, ovary cancer, malignant lymphomas, and lung adenocarcinomas.

Figure 2. Periodicity of tumor mutation rate within nucleosomes

(a) Schematic representation of alternating sequences of DNA with minor groove facing towards and away from histones.

(b) Observed and expected mutation rate of esophageal adenocarcinomas (top-left); relative increase of mutation rate (bottom-left) and its periodogram (right). Vertical broken lines represent stretches of minor groove facing away from histones.

(c) Signal-to-noise ratio (SNR; y-axis) of the strongest period (x-axis) in the relative increase of mutation rate of cohorts with phase 1 (top panel) or -1 (bottom panel). Cohorts with a relative increase of mutation rate with SNR>8 and q-value<0.05 are circled.

(d) Four examples of periodicity of the relative increase of mutation rate. Clockwise from top left corner: malignant lymphomas, ovary cancer, lung adenocarcinomas, skin melanomas.

Figure 3. Periodicity of the mutation rate within nucleosomes in individual tumors

(a) Distribution of minor-in relative increase of mutation rate across tumors of different cohorts. Tumors with significant SNR (q-value<0.1; SNR>8) of the relative increase of mutation rate appear colored in red.

(b) Tumors with significant minor-in relative increase of mutation rate across different cancer cohorts (colored following the legend in panel a). The contributions of different mutational signatures to the mutation landscape of selected tumors (and one normal skin cell) are highlighted in pie-charts below the graph. (Signatures with the largest contribution are indicated below each.) Bottom panels, relative increase of mutation rate signals of four selected samples.

Figure 4. Periodicity of the mutation rate within nucleosomes as a function of mutational signatures

(a) Signal-to-noise ratio (SNR; y-axis) of the strongest period (x-axis) in the relative increase of the rate of mutations contributed by different signatures with phase 1 (top panel) or -1 (bottom panel). Groups of mutations contributed by signatures with a relative increase of mutation rate with SNR>8 and q-value<0.05 are circled.

(b) Four examples of mutational signatures resulting in different patterns of relative increase of mutation rate periodicity. Clockwise from top left corner: signature 17, signature 1, signature 4, signature 7.

(c) SNR of the periodicity of the relative increase of the rate of mutations contributed by four signatures in nucleosomes with strong and weak rotational position (Methods).

Figure 5. Periodicity of DNA damage and repair and accessibility within nucleosomes
(a) Schematic summary of the approach to compute the relative increase of UV-induced damage (CPDs) within nucleosomes.
(b) Periodicity of the observed and expected CPDs (top left panel), the relative increase of the rate of CPDs (bottom left panel), the observed CPDs immediately after UV exposure and after 24 hours (top right panel), and the relative increase of the repair of CPDs (bottom right panel). Only CPDs at nucleosomes with strong rotational position were included in the computation of the SNR. The barplots represent the SNR of the periodicity of the relative increase of the rate of generation (bottom left panel) and repair (bottom right panel) of CPDs at nucleosomes with strong and weak rotational position.
(c) Periodicity of the observed and expected MeGs (top left panel), the relative increase of MeGs (bottom left panel), the observed MeGs immediately after MMS exposure and after 2 hours (top right panel), and the relative increase of MeGs repair (bottom right panel).
(d) Periodicity of the DNAse efficiency within human (left) and *S. cerevisiae* (right) nucleosomes. The barplots represent the SNR of the periodicity of the relative increase of accessibility across nucleosomes with strong and weak rotational position.

Figure 6. Periodicity of intra-species variation and inter-species divergence within nucleosomes
(a) Periodicity of rare variants in *H. sapiens* and *A. thaliana* populations. The barplots represent the SNR of the periodicity of the relative increase of the rate of rare variants across nucleosomes with strong and weak rotational position.
(b) Periodicity of the polarized C>T divergence between *H. sapiens* and *A. thaliana* and close species.

Figure 7. Eukaryotic DNA sequence periodicity
(a) WW periodicity of five eukaryotic species in stacked nucleosome-covered DNA.
(b) Schematic representation of the approach designed to compute the enrichment of WW periodicity in genomes.
(c) Enrichment of WW periodicity in several eukaryotic genomes.
(d) Schematic representation of the model linking the periodicity and the C→T divergence bias in the generation of *de novo* mutations with the development of the WW sequence periodicity in the genome.
(e) Simulation of the development of the WW sequence periodicity from a random DNA sequence through periodic biased mutations (top panel; gray area, confidence intervals computed from 1000 trials). The bottom panel illustrates the sequence periodicity observed in one randomly chosen trial.

Supplemental figures legends

Figure S1. Mutation rate periodicity between nucleosome-covered and linker DNA (related to Figure 1)

(a) Mutation rate periodicity computed across 3595 tumor cohorts analyzed not presented in Figure 1. For each cohort the three graphs that appear in Figure 1b for esophageal adenocarcinoma are presented. The top graph represents the observed and expected (with confidence intervals) mutation rate; the graph in the middle, the signal of the relative increase of mutation rate; and the bottom graph, its periodogram. The number of samples in the cohort (n), the strongest period in the periodogram (MP), its signal-to-noise ratio (SNR), and its associated p-value appear above the three graphs representing each cohort. The acronyms of the cohorts correspond to those in Table S1.
(b) Comparison of the SNR of the MP of the relative increase of mutation rate computed for each cohort included in the study employing (1) a penta-nucleotide context with mutation frequencies computed genome-wide (used for results reported in the paper, yellow dots); a tri-nucleotide context with mutation frequencies computed genome-wide (red dots); and a penta-nucleotide context with mutation frequencies computed from the genome after removal of nucleosome-covered DNA from MNAse data (see Methods for details, green dots). The latter is necessary to guarantee that the observed periodicity is not due to an artifact caused by different sequence composition of nucleosome-covered and linker DNA.

Figure S2. Mutation rate periodicity between minor-in and minor-out nucleosome-covered DNA stretches (related to Figure 2)
(a) Mutation rate periodicity computed across 3595 tumor cohorts analyzed not presented in Figure 2. For each cohort the three graphs that appear in Figure 2b for esophageal adenocarcinoma are presented. The top graph represents the observed and expected (with confidence intervals) mutation rate; the graph in the middle, the relative increase of the mutation rate; and the bottom graph, its periodogram. The number of samples in the cohort (n), the strongest period in the periodogram (MP), its signal-to-noise ratio (SNR), and its associated p-value appear above the three graphs representing each cohort. The acronyms of the cohorts correspond to those in Table S1.
(b) Comparison of the SNR of the MP of the relative increase of mutation rate computed for each cohort included in the study employing (1) a penta-nucleotide context with mutation frequencies computed genome-wide (used for results reported in the paper, yellow dots); a tri-nucleotide context with mutation frequencies computed genome-wide (red dots); and a penta-nucleotide context with mutation frequencies computed from the genome after removal of nucleosome-covered DNA from MNAse data (see Methods for details, green dots). The latter is necessary to guarantee that the observed periodicity is not due to an artifact caused by different sequence composition of nucleosome-covered and linker DNA.
(c) SNR of the periodicity of the relative increase of the rate of mutations at groups of nucleosomes with high (right bar in each cohort) and low (left bar) scores of rotational setting (see Methods).

Figure S3. Mutation rate periodicity between nucleosome-covered and linker DNA for mutations contributed by different sigantures (related to Figure 4)
(a) Correlation between the SNR of the MP of the relative increase of the mutation rate computed for sets of mutations obtained on the basis of two different approaches for signature decomposition (Sigfit and deconstructSig; see Methods). The high agreement between the two

methods guarantees that the use of the latter throughout the analyses accurately separates mutations based on their most likely contributing signature.
(b) Signal-to-noise ratio (SNR; y-axis) of the strongest period (x-axis) in the signal of the relative increase of the mutation rate within nucleosomes of groups of mutations contributed by different mutational processes. Dots representing cohorts with significant SNR appear encircled. Top panel: mutational signatures with relative increase of mutation rate with phase 1; bottom panel: mutational signatures with relative increase of mutation rate with phase -1 (see text).
(c) SNR of the relative increase of mutation rate (y-axis) of groups of mutations contributed by several mutational signatures (x-axis) in different cohorts. The SNR of the mutations in a given cohort of tumors contributed by a signature are represented as a circle. Groups of mutations with significant SNR (SNR>8; q-value<0.1) are colored following the legend of the corresponding tumor type, shown below the graph.

Figure S4. DNA UV-induced damage and repair (related to Figure 5)
(a) Periodicity within nucleosomes of (6-4)PPs generated in fibroblasts upon exposure to UV light. The two panels are analogous to those presented for CPDs (the other type of DNA UV-induced lesions) in the left side of Figure 5b.
(b) Periodicity within nucleosomes of CPDs (left), and (6-4)PPs (right) generated in the naked DNA of fibroblasts upon exposure to UV light. (The three graphs are analogous to those presented in panel a. The phase of the relative increase of both types of damage is the opposite (-1) to that observed for CPDs and (6-4)PPs formed across native DNA (i.e., containing nucleosomes).
(c) Periodicity of the repair of (6-4)PPs within nucleosomes of fibroblasts at 1h after UV irradiation. The two panels are analogous to those presented for CPDs (the other type of DNA UV-induced lesions) in the right side of Figure 5b.
(d) Comparison of the periodicity of the mutation rate in XPC wild-type (left) and XPC-mutant (tumors) both within nucleosomes (top panels) and in the nucleosome-linker alternation (bottom graph). The nucleosome-linker periodicity resulting from the impairment of NER at nucleosome-covered DNA is absent in XPC mutants.

Figure S5. SNR of sets of germline variants with varying minor allele frequency (related to Figure 6)
SNR of relative increase of the rate of rare (very lowly frequent; MAF<0.01), lowly frequent (0.01<=MAF<0.05) and frequent (MAF>0,05) germline variants detected across human populations within nucleosomes.

Figure S6. DNA sequence periodicity across eukaryotic genomes (related to Figure 7)
Enrichment of N bp WW periodicity in several eukaryotic genomes measured as the fraction of genomic chunks that have N bp as their maximum period and the odds-ratio of the N-bp period amongst all other integral periods between 6 and 19 across genomic chunks. N takes values between 6 bp and 19 bp (see Methods for details).

Supplemental Tables

Table S1. List of cohorts of whole-genome sequenced samples employed in the study (related to Figures 1,2,3,4)

Table S2. Metrics of periodicity (around and within nucleosomes) of somatic mutations grouped by cohorts and signatures (related to Figures 1,2,3,4)

Table S3. List of eukaryotic genomes employed to assess the WW periodicity of their sequence (related to Figure 7)

Table S4. Metrics of periodicity computed across integral periods between 6 and 19 for sevral eukaryotic genomes (related to Figure 7)

STAR Methods text

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Nuria López-Bigas (nuria.lopez@irbbarcelona.org)

METHOD DETAILS

Mapping nucleosomes

H. sapiens

The 147 bp length mid-fragments of high-coverage MNAse-seq reads (representing putative nucleosome dyads) mapped to the hg18 human genome assembly were obtained from (Gaffney et al., 2012) in wig format, and converted to the bed format using the wig2bed utility from BEDOPS (Neph et al., 2012). To find the most representative dyads, we first smoothed-out the mid-fragment counts using a 15-bp tri-weight kernel similar to that described in (Valouev et al., 2011). This method proceeds in two steps. First, we retrieved the raw dyad count per position, $d(i)$, and smoothed-out the resulting signal using the kernel

$$K(x) = \left(1 - \left(\frac{x}{15}\right)^2\right)^3 \cdot 1_{|x|<15},$$

thereby producing the kernel-smoothed dyad count

$$D(i) = \sum_{j=1}^{N} K(i-j) \cdot d(j),$$

where N is the length of the sequence. Second, at each position i, we correct $D(i)$ dividing by an approximation of the total number of counts in the $i \pm 150$-bp interval, resulting in the following stringency metric:

$$S(i) = \frac{D(i)}{\sum_{j=i-150}^{i+150} \alpha \cdot \frac{1}{15} D(j),} \text{ with } 1/\alpha = \int_{-1}^{1} (1-u^2)^3 du.$$

The resulting bed files with the smoothed dyads of each chromosome were merged and converted into a wig file using the BedgraphToBigWig tool (Kent et al., 2010). The *local maxima* of the smoothed counts, which represent the highest fraction of "well positioned" nucleosomes covering a position, were obtained using bwtool (Pohl and

Beato, 2014) with the parameters “find local-extrema -maxima -min-sep=150”. The mid-fragment with the highest number of reads within a 30 bp interval was selected to contain the putative and most representative dyad, and any other mid-fragment in that interval was discarded. In case of a tie between two or more mid-fragments, the dyad closest to each *local maxima* was selected. Dyad coordinates were lifted over from hg18 to hg19 using CrossMap (Zhao et al., 2014). We extended the sequences around dyads by 73 bp at each side, and we kept the resulting nucleosome-covered sequences only if all their nucleotides were mappable according to the CRG36 Alignability track.

To obtain the set of intergenic nucleosomes, we retrieved the coordinates of genic regions from Gencode (Harrow et al., 2012), extended them by 500 bp on each side of their start and end boundaries, and removed all nucleosomes overlapping the extended regions. The final human dataset of intergenic nucleosomes comprised 3,759,105 instances. We call this subset the human nucleosome set.

We also created a set of human genic nucleosomes, which we used to analyze exonic mutations from exome sequencing data. To this end, we took the sequences extending 73-bp on each side of each dyad, and intersected them with exon coordinates from Gencode. We removed any nucleosome with nucleotides overlapping introns or the first or last exons of genes.

A. thaliana

Paired-end reads from high-coverage MNAse-seq experiment from flowers of *Arabidopsis thaliana* were obtained from (Zhang et al., 2015) using NCBI’s fastq-dump with the commands “fastq-dump --split-3 -F --read-filter pass”. Reads were mapped to the TAIR10 genome using bowtie (Langmead et al., 2009) with the following parameters: “bowtie -q --nomaqround --phred33-quals -S --chunkmbs 200 --seed 123”, and the resulted bam files were processed using samtools (Li et al., 2009). Mapped fragments of 146-148bp length were selected. The coordinates of the set of intergenic nucleosomes were finally chosen applying the same procedure as explained for *H.*

sapiens (i.e., kernel-smoothing, selection of representative dyads, removal of genic regions).

S. cerevisiae

The genomic coordinates of *S. cerevisiae* nucleosome dyads obtaind using a chemical-mapping approach in the saccer2 genome were obtained from (Brogaard et al., 2012). Dyads mapped at distances shorter than 147bp were removed from further analysis. The dyads were lift over to saccer3 using CrossMap. These dyads were intersected with the coordinates of genes to obtain the set of intergenic nucleosomes.

D. melanogaster

The coordinates of intergenic *Drosophila melanogaster* nucleosomes mapped using MNase-seq were directly obtained from the supplementary material of (Langley et al., 2014) and filtered using the same protocol described in the original analysis.

M. musculus

The genomic coordinates of non-overlapping *M. musculus* nucleosome dyads obtained using a chemical-mapping procedure in mouse ES cells were downloaded from (Voong et al., 2017). Dyads falling within unmappable regions (according to the mouse CRG36 Alignability track) and the ones falling in genic regions (extended 500bp on each side) were removed. The sequences of nucleosome-covered DNA were obtained selecting the regions flanking the dyads as explained above.

Segments of DNA with minor groove facing to or away from nucleosomes

The positions relative to the dyad of the two nucleotides at the center of the stretches of DNA with the minor groove facing the histones and away from them were obtained from (Cui and Zhurkin, 2010). We use these positions (i) as reference for the vertical lines of all 117-bp wide figures; (ii) to color minor-in and minor-out segments in the signals of relative increase of the rate of mutations, DNA damage, repair, polymorphisms, and C→T polarized divergence (by extending the DNA sequence around each of them, except for the center of the dyad and the 10bp flanking it); (iii) as reference to select the

nucleotides to compute the minor-in relative increase of mutation rate shown in Figure 3 (see below); (iv) as reference to compute the phase of relative increase of the rate of mutations, DNA damage, repair, polymorphisms, and C→T polarized divergence (see below); and (v) as reference to compute the rotational score of nucleosomes (see next section).

Classifying nucleosomes based on the strength of their rotational setting

The rotational positioning of nucleosomes is maintained if the histones core moves preferentially in 10 bp intervals along the DNA, so that the same minor groove stretches preferentially face it or away from it. To quantify the strength of the rotational positioning of nucleosomes, for each dyad in our human nucleosome set we computed a "rotational score" RS as the fraction of the total of reads mapping the nucleosome region that fall in positions of the DNA with the minor groove facing the histones (R_{out}), that is, $RS = R_{out}/R_{total}$. To this end, we used as reference the minor-in positions described in the previous section. The set of nucleosomes with high score of rotational setting (employed in several SNR comparisons) was interated by all nucleosomes with *RS*=1. We then ranked all remaining nucleosomes by ascending *RS*, and selected from the top ranking the same number of nucleosomes as in the set of high rotational setting to integrate the set of nucleosomes with low score of rotational setting. The SNR of the periodicity of the rates of mutations, DNA damage, repair and polymorphisms within these two sets of human nucleosomes were compared (Fig. 4, 5 and 6). Only the nucleosomes in the set with high score of rotational setting of *H. sapeins* and *A. thaliana* were used in the analysis of the periodicity of the C→T polarized divergence.

Somatic mutations

Whole-genome somatic mutations identified in 5766 tumors of 59 cohorts sequenced by ICGC projects and 505 tumors across 14 cohorts sequenced by TCGA were obtained from the ICGC data portal and a publication (Fredriksson et al., 2014; ICGC, 2010), respectively. The name of each cohort and the country charged with the corresponding sequencing project were appended to the names of tumors sequenced by ICGC (http://docs.icgc.org/submission/projects/). The same nomenclature (except for the

country) was followed for tumors sequenced by TCGA, when possible. For cohort-based analyses, in cases of a tumor type represented by two cohorts we used the cohort with more mutations. Only PASS-labeled somatic variants were analyzed. We filtered out all mutations falling outside mappable regions of the genome according to the CR36 mappability track. Finally, cohorts with fewer than 500 mutations mapping to intergenic nucleosomes (see beolw) were discarded. After these filtering steps, we obtained 28 tumor cohorts comprising 3494 tumors bearing 60,152,954 SNVs (Table S1). Filtered whole-exome somatic mutations were obtained from the repository of the TCGA PanCanAtlas initiative (Ellrott et al., 2018). Whole-genome somatic mutations identified in one normal eyelid skin sample were obtained from (Martincorena et al., 2015). Whole-genome somatic mutations detected in 8 XPC wild-type and 5 XPC-mutant tumors were obtained from (Zheng et al., 2014). The same mappability filter was applied to all these sets of mutations.

Spectral analysis of the signals

Spectral analysis seeks to deconstruct the periodic components of a signal. In order to assess the degree of periodicity of a given signal S, we resort to the computation of its power spectrum. Signals in our context constitute discrete functions in the DNA sequence position domain which map each position to a magnitude of interest, e.g., mutation rate, repair rate, di-nucleotide motif autocorrelation, and the relative increase of mutation rate, DNA damage, repair, polymorphisms, and C→T polarized divergence. Therefore periods will be given in base-pairs (bp).

Power Spectrum

The power spectrum S^* of a discrete-time signal $S = \{S_n\}$ with mean zero is the Discrete Time Fourier Transform (DTFT) of its covariance vector $r_k = E(S_{n-k} \cdot S_n)$ (Stoica and Moses, 2004). When S is given by a finite sample $\{S_n\}$of length N, S^* can be approximated as:

$$S^*(P) \approx \frac{1}{N}\left|\sum_{n=1}^{N} S_n \cdot exp\left(\frac{-2\pi i n}{P}\right)\right|^2.$$

For convenience we chose to define S^* in the period domain, and we refer to its graph as "periodogram". Intuitively, S^* can be interpreted as a continuous signal that maps each period P to a magnitude encoding the contribution of the periodic component of period P.

Signal-to-Noise Ratio

A standard approach to interpret the power spectrum consists in computing a signal-to-noise ratio (SNR) representing the relative amount of power at a certain period compared to a baseline overall contribution by other periods. A value of SNR provides a systematic measurement of how much periodicity a signal has and which periods contribute the most to this periodicity. We define the SNR centered at a period P as the ratio of the maximum power attained within a specified period interval $I(P) = \{P - \delta \leq x \leq P + \delta\}$ and the median power throughout the complement of $I(P)$, i.e.,

$$SNR(P) = \frac{max\{S^*(q), q \in I(P)\}}{median\{S^*(q), q \notin I(P)\}}.$$

Unless otherwise specified, we let $I(P)$ to be 1 bp in diameter, i.e., $\delta = 0.5$ bp. Thus defined, the SNR measures to what extent the signal can be explained by a single periodicity in the vicinity of the center P.

Normalized Power Spectrum

The SNR value does not depend on the linear scale in which the values of S^* are given. Therefore, as most analyses concerned in this work imply the computation of the SNR and the Maximum Power Period (MP), both linear scale invariant, we will render a "normalized power spectrum" expressed in arbitrary units (a.u.). Given a period interval of interest $L = \{P \leq x \leq Q\}$, the normalized power spectrum, denoted S_u^*, will be a re-scaling of S^* so that its mean through L is 1. Since we encode S^* as a discrete function defined in a grid of L, we can compute this normalization as:

$$S_u^*(p) = (Q - P + 1) \cdot \left(\sum_{j=P}^{Q} S^*(j)\right)^{-1} \cdot S^*(p)$$

Phase Shift Analysis

The phase shift between two signals S_1 and S_2 at a given period P, denoted $\Delta_P(S_1,S_2)$ is defined as the unique value $-\pi \leq v < \pi$ such that $\theta_1(P) - \theta_2(P) - v$ is a multiple of 2π, where θ_1 and θ_2 denote the phase functions of the DTFTs of S_1 and S_2, respectively. To describe whether a periodic component of S is more likely in phase rather than out of phase with respect to a reference sinusoidal signal R at period P, we establish the following definition:

> if $|\Delta_P(S,R)| \leq \pi/2$, we denote the phase of S as 1 relative to P; otherwise, we denote its phase as -1.

Stacked sequence analyses

We extended the nucleotide context of each mutation in a cohort by 2 bp on each side. For each context, the relative frequency of each type of mutation was calculated by dividing the number of mutations with that context by the abundance of the context in the mappable intergenic genome (defined as described above). We call this array of frequencies the Extended Mutation Spectrum (EMS).

The sequence of each dyad-centered nucleosome was extended by 1000 bp (for nucleosome/linker analyses) or 58 bp (for minor-in/minor-out analyses) on each side. Mutations were then intersected with these 2001- or 117-bp sequences using BedTools (Quinlan and Hall, 2010), and thus their absolute genomic positions were transformed to positions relative to the dyad. We then assessed the likelihood of observing a mutation at each position in each of these dyad-centered sequences in two different ways.

Expected Count by Frequency

For each position *p* within each sequence we obtained the 5-mer nucleotide context. Then, using the EMS as a null mutational model, we weighted p by its respective context-specific frequency, and made the sum of weights across all n=2001 or n=117 values equal to 1. In other words, let $f = (f_c)$ be the EMS and $c(p)$ be the pentanucleotide context at the *p-th* position, the vector of weights $w = (w_p)$ across the specific 2001- or 117-bp sequence is given by:

$$w_p = f_{c(p)} / \sum_{q=1}^{n} f_{c(q)}.$$

We then stacked all 2001- or 117-bp sequences and copmuted the expected count at position p of the stack as $M \cdot w_p$, where M is the count of mutations intersecting the sequence where p is located. Finally, the expected count at a given position p of the stack of aligned sequences, denoted $E(p)$, is obtained as the sum of all the expected counts at sequence positions mapping to p. Then, for each position of the stack we computed the fold increase of the observed mutation count mapping to p, denoted $O(p)$, with respect to $E(p)$, which we termed relative increase of mutation rate:

$$Relativeincrease(p) = \frac{O(p) - E(p)}{E(p)}.$$

We repeated all calculations of the expected mutation rate, and subsequently of the relative increase of the mutation rate and its SNR using the probabilities of tri-nucleotide changes instead of penta-nucleotides. The results, shown in Figures S1b and S2b were similar to those obtained with penta-nucleotide changes. The same approach was used to compute the relative increase of the DNA damage, DNA repair, germline genetic variants and C→T polarized divergence (see below).

Expected Count by Randomization

The second method consists in randomly placing the number of mutations observed in each sequence with a context-dependent position probability that is proportional to the vector of weights w corresponding to the sequence, and repeat this process 1000 times. We are then able to compute the frequency of these randomly placed mutations at each position of the 2001- or 117-bp stack of sequences. We use this frequency to compute the confidence intervals of the expected mutation rate represented in Figures 1 and 2.

Periodicity of the relative increase of mutation rate

To identify the periodic components of the discrete one-dimensional signal defined by the relative increase of mutation rate (or DNA damage, repair, germline genetic

variants, and C→T polarized divergence), we computed its power spectrum as described above. In order to rule out oscillations with smaller periods (noise), our target signal is obtained upon smoothing the relative increase of mutation rate with a cubic spline (R/stats/smooth.spline).

For each randomization of the mutations in the stack of aligned sequences computed as explained above, we can derive a random relative increase of mutation rate:

$$RelativeIncrease_r(p) = \frac{R(p)-E(p)}{E(p)},$$

where $R(p)$ is the sum of randomized mutation counts mapping to position p and $E(p)$ is the expected count of mutations as in the "Expected Count by Frequency" approach. We compute the $Relativeincrease_r$ for the 1000 randomizations. Upon computing the SNR for the relative increase of mutation rate and $Relativeincrease_r$ signals, we can compute an empirical p-value of the relative increase of mutation rate by asking how often the SNR calculated from the set of $Relativeincrease_r$'s is greater than the SNR value obtained from the relative increase of mutation rate. This empirical p-value is the one shown in Figures 1, 2, 3, 4. Multiple test correction was carried out using Benjamini-Hochberg FDR yielding a q-value for each cohort. The threshold of q-value and SNR used to deem a relative increase of the rate of mutations significantly periodic is indicated in each of these figures.

In order to plot the nucleosome-linker mutation rate and relative increase of mutation rate signals, we extended sequences by 1000 bp at each side of every dyad and then calculated the relative positions of neighbouring dyads. After aggregating all the distances, a cubic spline smoothing was applied (R/stats/smooth.spline) and the local maxima were selected. From each local maxima, we labeled "nucleosome" the range between ± 73 bp at each side of the local maxima. The sequence of segments between nucleosomes was deemed to correspond to linkers.

The signals of relative increase of mutation rate, damage, repair, polymorphisms, and C→T polarized divergence were smoothed using the R/stats/smooth.spline function and plotted according to whether they are nucleosome/linker or minor-in/minor-out.

Mutational signatures fitting and assignment

We assessed which of the 30 COSMIC mutational signatures were active in each tumor sample in the uniformly processed TCGA 505 cohort (Fredriksson et al., 2014) by using the deconstructSigs (Rosenthal et al., 2016) R package. The resulting weights (for each of the signatures active in each sample) were then multiplied by the total number of mutations in that sample to compute the signature exposure (i.e., the number of mutations potentially contributed by each signature). Then, using the exposure and the corresponding mutational signatures, we computed the probability for each mutation to be contributed by the different signatures identified in the sample). Finally, for each mutation, the signature with the maximum probability was considered as the one contributing it (Morganella et al., 2016), and all mutations coming from the same signature were aggregated. The same procedure was used in PanCanAtlas cohorts, with the parameter 'exome2genome' as a normalization method.

As a control, we performed the same task using another package called SigFit (Gori and Baez-Ortega, 2018), which runs a signature fitting method based on a Markov Chain Monte Carlo (MCMC) sampling to fit the input set of mutational signatures to a mutational catalogue, using a Non-Negative Matrix Factorization (NMF) model to sample from. We applied the method with the parameters iter=13000, warmup=3000, and hpd_prob=0.90 to get the exposures and 90% highest posterior density intervals. The comparison between the periodicity of the relative increase of mutation rate in signatures obtained from the two assignments is shown in Supplementary Figure 3A.

DeconstructSigs was also run on each ICGC and TCGA cohort independently to produce Figure S3c**.**

DNA damage data

UV damage

The genomic positions of CPDs and (6-4)PPs captured and mapped using the Damage-seq technique were obtained from (Hu et al., 2017). We located the start of each

pyrimidine dimer (taking into account the strand to which the read was mapped) and extended 2 bp to each side to assess the frequency of formation of each type of lesion within all possible penta-nucleotide sequence contexts as explained above for mutations.

MMS damage

The genomic positions of MMS-methylated bases in saccer3 genome version were downloaded from (Mao et al., 2017). As recommended by the authors, in order to explore the efficiency of BER only reads supporting methyl-guanines were selected. The site of alkylation was extended 2bp at each side to assess its frequency within all possible penta-nucleotide sequence contexts as explained above for mutations.

Repair inference

For each position p of the nucleosome stack, we compute the repair rate as:

$$RR(p) = \frac{Di(p) - Df(p)}{Di(p)},$$

where $Di(p)$ and $Df(p)$ denote the total amount of damage in the full stack of nucleosomes observed at the initial (0h) and final (24h for CPD, 1h for PP 6-4, 2h for MMS) time-points of the experiment, respectively. Therefore, if all the amount of damage at a position of the stack at the initial time-point is missing at the final time-point, $RR(p) = 1$; whereas if the total amount of damage does not change, $RR(p) = 0$.

Accessibility

Sequencing reads of fragments of DNA obtained by digestion with DNAse (DNAseq) of the human and yeast genomes were obtained from (Degner et al., 2012) and (Zhong et al., 2016) respectively. In order to account for the DNAse nick cut preferences, the start position of each read was extended 2 bp to both sides to get the 5-mer-based expected distribution count as explained above for mutations.

Germline variants

Rare germline variants detected within *H. sapiens* and *A. thaliana* populations were obtained from the 1000 genomes phase 3 (Gibbs et al., 2015) and the 1001 Genomes Project (Alonso-Blanco et al., 2016), respectively. Only SNVs with PASS filter and with allele frequency<0.01 were kept. Variants obtained from the 1001 Genomes Project data was processed using vcftools (Danecek et al., 2011).

Ancestral states

Human ancestral state

In order to reconstruct the genome of the most recent common ancestor of *H. sapiens* and *P. troglodytes*, a procedure similar to that described by (Langley et al., 2014) was employed. First, the UCSC 20-multi-way multiple alignment of 19 mammalian (16 primates) with human was obtained (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz20way/). The closest species to human selected was *Pan troglodytes,* while *Gorilla gorilla* was chosen as an outgroup in the genome sequence comparison. The polarized C>T divergence in the human lineage consists of sites that are conserved between these two species, but have diverged in the human lineage. On detail, given the alignment between *H. sapiens*, *P. troglodytes* and *G. gorilla*, we consider the following ad-hoc definitions: a “polarized site” is a site that: i) is conserved in the *Pan troglodytes* and *Gorilla gorilla* genomes; ii) has conserved 3’ and 5’ flanks across the three species; iii) is C or a G. Further, we say that a polarized site is a “divergent site” if the in the human genome we observe a T (in case of ancestral C) or A (in case of ancestral G) at the corresponding position.

We intersected the polarized divergent sites with the 117-bp wide sequences representing nucleosomes. We computed the per-position proportion of polarized sites that are divergent to render a discrete signal defined on each nucleosome position relative to the dyad. Finally we computed the power spectrum of the signal to assert its periodic structure.

To define the ancestral and polarized divergent sites in the genome of *Arabidopsis thaliana* with respect to those of *A. lyrata,* and *B. rapa*, the same approach was

followed. In this case, the whole-genome alignments were downloaded from ENSEMBL (ftp://ftp.ensemblgenomes.org/pub/current/plants/maf/). Only the set of *Arabidopsis* nucleosomes with high score of rotational setting (see above) was used in this case.

Genome-wide periodicity analysis

Genomic sequences of eukaryotic organisms

DNA sequences of the eukaryotic genomes displayed in Figures 7c S6 (Table S2) were retrieved as FASTA files from the UCSC Genome Browser (Fujita et al., 2011) and ENSEMBL Genome Browser (Zerbino et al., 2018). Repetitive regions were masked so we can rule them out from our categorical periodicity analysis.

General approach to compute the WW periodicity

We devised and implemented a method to assess the periodicity of the WW motif {AA, AT, TA, TT} in the categorical signal defined by an input DNA sequence. Moreover, we devised a statistical analysis to compare the strengths of the signal of WW periodicity across genomes. The analysis comprises two main steps: i) encoding the categorical signal defined by the motif of interest in the DNA sequence as an autocorrelation continuous function; ii) performing the spectral analysis using this continuous function as input. The methodology extensively resorts to previous work examining the periodic structure of categorical signals arising from DNA motifs; see, for instance, (Herzel et al., 1999; Mrázek, 2010).

Note that this part of our analysis is not equivalent to the analysis of motif frequencies at each position of a stack of aligned DNA sequences; instead, we are making the more general inquiry of whether the WW periodicity arises in the categorical signal defined by long stretches of DNA regardless of any alignment.

Chunk Preparation

Given the masked (see below) chromosome DNA sequences as input, we first created a set of 1 Mb chunks covering the chromosomes by sequentially taking 1 Mb sequences from the starting position of the chromosome and shifting 0.5 Mbp downstream at a

time. As a result, each position of the genome is covered by two chunks, except the 500 nucleotides at chromosome ends. Chunks shorter than 1 Mb (remains of chromosome split) were excluded from the analysis. In this step we used the UCSC program faSplit.

For each chunk we recorded the number of motif pairs at a specific distance *d* for $2 \leq d \leq 120$ (bp). For each chunk we also recorded the probability of observing a nucleotide, say z, conditioned to its 5' flanking dimer, say xy, for $x, y, z \in \{A, C, G, T\}$. This information can be encoded as a 16×16 transition probability matrix M defining a 16-state Markov chain:

$$M = \left(a_{xy,zt} | x, y, z, t \in \{A, C, G, T\}\right),$$

such that $a_{xy,zt} = 0$ provided $y \neq z$. For each chunk we set up a stochastic chunk generator which draws from this model by iteratively adding nucleotides based on the last dimer added –the initial dimer drawn from the stationary distribution of the model. In sum, for each observed chunk we derive a second-order 16-state Markov chain fitting its di-nucleotide composition and we draw a null sample of 1 Mb chunks out of it. We generated 100 random 1 Mb samples for each chunk, that we later used to derive statistics that let us control for periodicity effects that may arise by di-nucleotide content alone.

WW motif encoding

For each chunk, using the counts of WW pairs at each distance $2 \leq d \leq 120$(bp) and its di-nucleotide content, we derive an autocorrelation score $A(d)$ that is defined for each distance as the ratio of two values: i) the observed frequency of WW pairs at distance d, i.e., $f_{WW}(d) = N_{WW}(d)/(N - d)$, where $N_{WW}(d)$ is the count of WW pairs at distance d and N is the length of the chunk; ii) the expected frequency of WW pairs at distance d, estimated as $f_{WW}^2(0)$, where $f_{WW}(0)$ is the frequency of WW. Therefore we get,

$$A(d) = \frac{f_{WW}(d)}{f^2{}_{WW}(0) \cdot (N - d)}.$$

A putative WW periodicity of period P in the DNA sequence is expected to be revealed in the $A(d)$ curve as a periodic component period P, with local maxima at multiples of P.

To carry out the spectral analysis of $A(d)$ to this end, we first took several important pre-processing steps.

First, for smaller values of d, the curve $A(d)$ may show a very strong effect due to short range interactions that may not have a direct connection with the periodicity we aim to study. For higher values of of d, the curve $A(d)$ may quickly decay in amplitude. Then in our analysis we committed to study $A(d)$ in a restricted interval, trimming the signal below 30 bp and above 100 bp to avoid confounders. The choice of trimming thresholds was based on the analysis described in (Mrázek, 2010).

Second, the scientific literature consistently reports the existence of a pervasive 3 bp periodicity component in DNA sequences from genomes comprising eubacteria, archaea and eukarya domains, which has been associated to protein coding DNA (Fickett and Tung, 1992). In order to rule out this effect from our analysis, we smoothed the trimmed $A(d)$ curve with a 3-bp moving average, yielding $\hat{A}(d)$.

Third, $\hat{A}(d)$ curves in general show a clear non-linear mean decay; in order to enable the computation of the power spectrum we had to render the zero mean input signal that best represented the spectral structure in the period scale we focus on. Therefore we proceeded to de-trend the signal. In other words, we computed the quadratic least squares fit $Q(d)$ of $\hat{A}(d)$ yielding $S(d) = \hat{A}(d) - Q(d)$.

After trimming, 3-bp smoothing and de-trending, we obtained a target signal for the subsequent spectral analysis, which we denoted S.

Spectral Analysis of Chunks

Each 1 Mb chunk and the corresponding 100 randomizations therefore yielded a continuous signal S that we could further analyze. Thereafter, we computed the power spectrum, as described above. For comparison of the power spectrum of chunks with the corresponding randomizations, we required that the power spectrum was not given in normalized form.

Periodicity Statistics

To understand what is the relevance of the 10 bp periodicity across eukaryotic genomes compared to other periodicities, we limited our study to a reduced set of periods: the integral periods from 6 bp up to 19 bp. To this end we computed the power spectrum of S only in the period interval between 5 bp and 20 bp with a resolution of 100 points. For each 1 Mb chunk and each period P we computed the median fold-power increase (FPI) and an associated empirical p-value:

$$FPI(P) = \frac{S^*(P) - m(P)}{m(P)},$$

where $m(P)$ is the median power at P across the 100 randomizations of the chunk.

For each chunk we also computed the period at maximum power (MP) and the signal-to-noise ratio (SNR) centered at all periods of the reduced set. To answer the question of whether a particular genome exhibits a detectable WW periodicity of 10 bp, we counted the number of its chunks that exhibit a significant FPI at each period between 6 bp and 19 bp. An FPI was deemed significant if its associated empirical p-value was smaller than 0.01. The enrichment of chunks that have significant FPI at 10 bp was given as an odds-ratio based on the count of chunks that have significantly high FPI, at each period: the observed odds to draw a significant FPI at 10 bp divided by the expected odds, which assumes a uniform distribution across periods:

$$\mathrm{OR} = \frac{o_{10}/(N-o_{10})}{e_{10}/(N-e_{10})},$$

where o_{10} and e_{10} are the observed expected number of chunks wiht significant FPI at 10 bp and N is total number of times any chunk had significant FPI at any period in the reduced set. We also computed a p-value usign a G-test (Sokal R., 1995):

$$g = 2 \cdot \left[o_{10} \cdot log \frac{o_{10}}{e_{10}} + (N - o_{10}) \cdot log \frac{N - o_{10}}{N - e_{10}} \right] \sim \chi_1^2.$$

For each periodicity P in the reduced set under study we computed the proportion of chunks that have maximum power within the interval $I = \{P - 0.5 \leq x \leq P + 0.5\}$.

Repeating this approach for all the eukaryotic genomes under study (Table S3), we obtained for each, the WW 10-bp periodicity power enrichment (odds-ratio), an associated p-value and q-value (FDR) and a proportion of chunks showing maximum power peak about 10 bp (Fig. 7c). To assess the relevance of other periodic structures of the same motif, we conducted the same analysis centered at the other integral periods between 6 bp to 19 bp (Fig. S6).

Sequence periodicity simulator

In order to investigate the emergence of periodic structures in the signal defined by the proportion of WW di-nucleotides for each position in the stack of aligned nucleosomal DNA, we define an operational model of differential mutagenesis.

Given a 1 Mb DNA sequence covered by nucleosomes and linkers, the model defines the stochastic rules to introduce new single nucleotide substitutions, thereby producing a new sequence: each such step is termed a “generation”. A key ingredient of the model consists in modulating the probability given by the sequence content with a vector of weights that depends on the position relative to the closest dyad. This procedure can be iterated and repeated with several starting chunks so that we can study how the sequences are shaped over the generations in terms of the frequency of WW motifs in the stacked nucleosome-covered DNA sequences.

In each generation a constant number of mutations M is randomly introduced in the sequence. The positions that undergo mutation are drawn without replacement from a probability distribution that depends on: i) a 6-channel, strand-symmetric mutational spectrum S corresponding to the human spontaneous mutations described in (Long et al., 2018), and ii) a weight function $W(x)$ that modulates the probability given by the mutational spectrum and only depends on the distance from x to the closest dyad,

denoted y. If $b(x)$ denotes the base at x, then the probability $Pr(*|\mathrm{x})$ that a mutation occurs at position x is defined as:

$$Pr(*|\mathrm{x}) \propto S(*|\mathrm{b(x)}) \cdot W(x),$$

where

$$W(x) \propto 1 - A \cdot cos\left(\frac{2\pi \cdot (x-y)}{P}\right).$$

In particular, $W(x)$ is periodic with period P. The weight function $W(x)$ is a discrete sinusoidal wave with local maxima at minor-in positions and a relative amplitude of A with respect to the mean. For our model we set $P = 10.3$ (bp) and $A = 0.04$, that is, we set a conservative relative amplitude with respect to the mean of 4%.

To run one simulation, we initialize with a randomly generated 1Mb chunk with uniform nucleotide probability. We run the model for 6000 generations and 100 mutations per generation. Every 100 generations we compute the WW frequency per position of the aligned nucleosomes and carry out the spectral analysis of the resulting signal. From the periodogram, we obtain the Maximum Power Period (MP) and the SNR centered at MP (which are represented in Fig. 7e).

QUANTIFICATION AND STATISTICAL ANALYSIS

Our paper consists of statistical analyses applied to somatic mutations, germline variants, interspecies polarized divergence, DNA damage, DNA repair and DNA sequence periodicity across eukaryotes. All these analyses, many of which were designed and implemented by us in Python, R and GNU bash scripts using interfaces such as Ipython (Perez and Granger, 2007) and readily available libraries such as pandas (McKinney, 2010) and numpy (Oliphant, 2006) are described at length in the previous sections of STAR Methods. Details and results of these analysis (the number of samples, mutations or polymorphisms, maximum power period of signals, the signal-to-noise ratio, and the associated p-values) are indicated for a few examples in pertinent Results sections and within the main Figures. The details and results of all analyses carried out (both present and absent from main Figures) are presented in

Supplementary Figures and Tables. Figures were constructed using Matplotlib (Hunter, 2007).

DATA AND SOFTWARE AVAILABILITY

All software produced by the study (including scripts needed to reproduce all results described in the paper) is available at https://bitbucket.org/bbglab/nucleosome-periodicity.

Figure 1

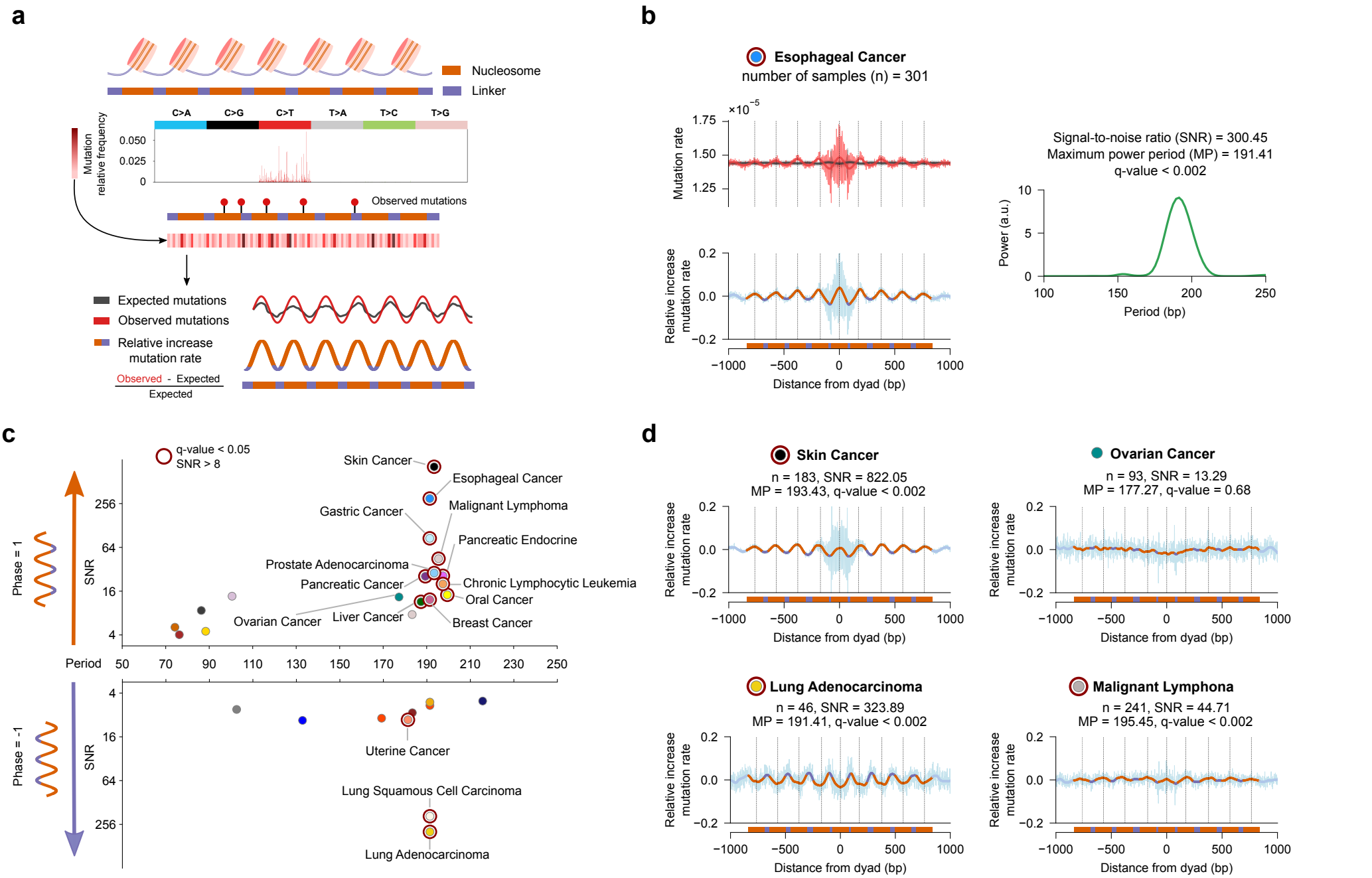

a
Nucleosome
Linker
C>A
C>G
C>T
T>A
T>C
T>G
Mutation relative frequency
0.050
0.025
0
Observed mutations
Expected mutations
Observed mutations
Relative increase mutation rate
Observed - Expected
Expected
b
Esophageal Cancer
number of samples (n) = 301
×10−5
Mutation rate
Relative increase mutation rate
Distance from dyad (bp)
Signal-to-noise ratio (SNR) = 300.45
Maximum power period (MP) = 191.41
q-value < 0.002
Power (a.u.)
Period (bp)
c
q-value < 0.05
SNR > 8
Phase = 1
SNR
Period
Skin Cancer
Esophageal Cancer
Malignant Lymphoma
Gastric Cancer
Pancreatic Endocrine
Prostate Adenocarcinoma
Pancreatic Cancer
Chronic Lymphocytic Leukemia
Oral Cancer
Ovarian Cancer
Liver Cancer
Breast Cancer
Phase = -1
Uterine Cancer
Lung Squamous Cell Carcinoma
Lung Adenocarcinoma
d
Skin Cancer
n = 183, SNR = 822.05
MP = 193.43, q-value < 0.002
Ovarian Cancer
n = 93, SNR = 13.29
MP = 177.27, q-value = 0.68
Lung Adenocarcinoma
n = 46, SNR = 323.89
MP = 191.41, q-value < 0.002
Malignant Lymphona
n = 241, SNR = 44.71
MP = 195.45, q-value < 0.002

Figure 2

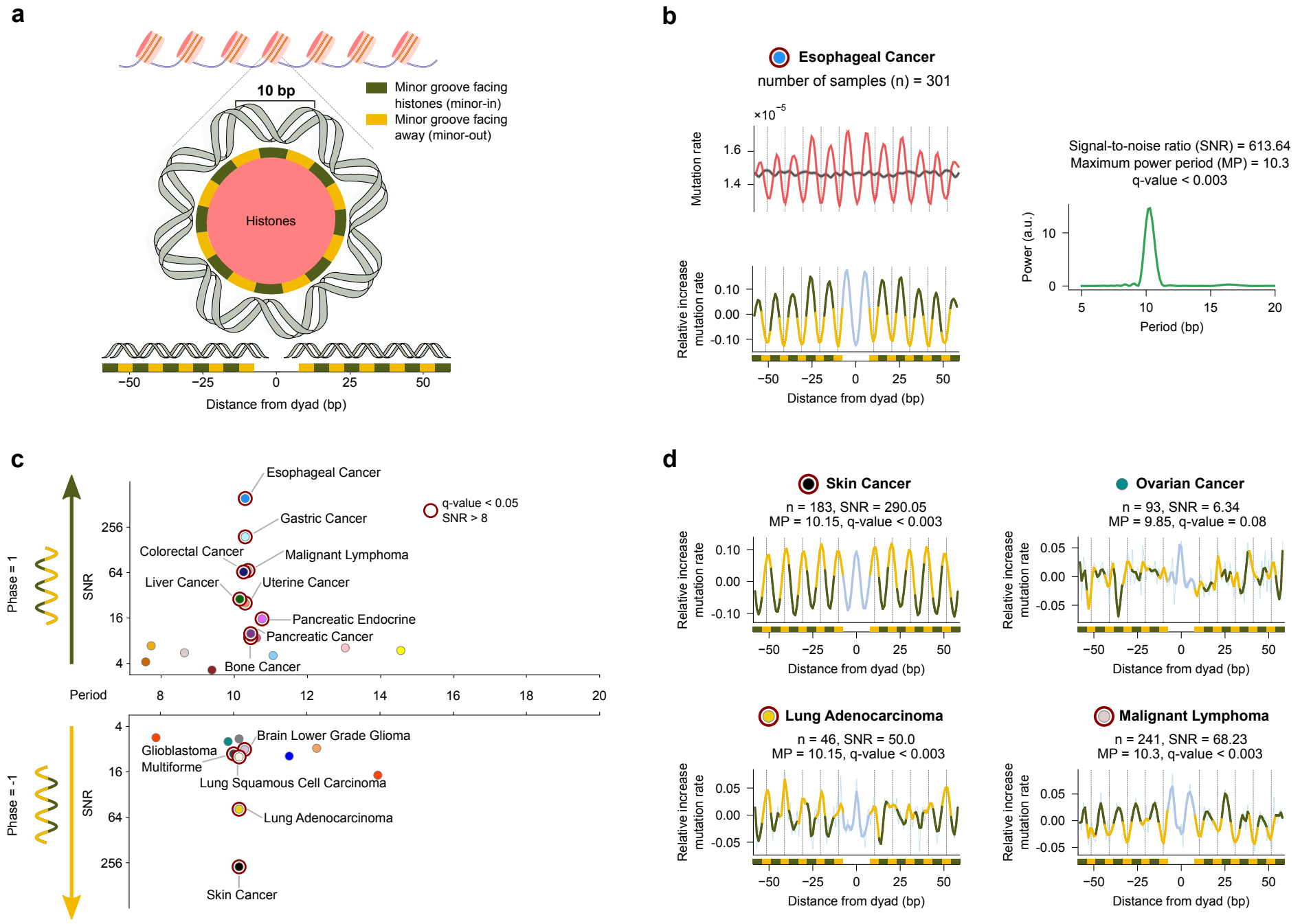

a
10 bp
Minor groove facing histones (minor-in)
Minor groove facing away (minor-out)
Histones
−50 −25 0 25 50
Distance from dyad (bp)
b
Esophageal Cancer
number of samples (n) = 301
×10^−5
Mutation rate
1.6
1.4
Relative increase mutation rate
0.10
0.00
-0.10
−50 −25 0 25 50
Distance from dyad (bp)
Signal-to-noise ratio (SNR) = 613.64
Maximum power period (MP) = 10.3
q-value < 0.003
Power (a.u.)
10
0
5 10 15 20
Period (bp)
c
Phase = 1
SNR
256
64
16
4
Esophageal Cancer
Gastric Cancer
Colorectal Cancer
Malignant Lymphoma
Liver Cancer
Uterine Cancer
Pancreatic Endocrine
Pancreatic Cancer
Bone Cancer
q-value < 0.05
SNR > 8
Period
8 10 12 14 16 18 20
Phase = -1
SNR
4
16
64
256
Brain Lower Grade Glioma
Glioblastoma Multiforme
Lung Squamous Cell Carcinoma
Lung Adenocarcinoma
Skin Cancer
d
Skin Cancer
n = 183, SNR = 290.05
MP = 10.15, q-value < 0.003
Ovarian Cancer
n = 93, SNR = 6.34
MP = 9.85, q-value = 0.08
Lung Adenocarcinoma
n = 46, SNR = 50.0
MP = 10.15, q-value < 0.003
Malignant Lymphoma
n = 241, SNR = 68.23
MP = 10.3, q-value < 0.003
Relative increase mutation rate
Distance from dyad (bp)
−50 −25 0 25 50

Figure 3

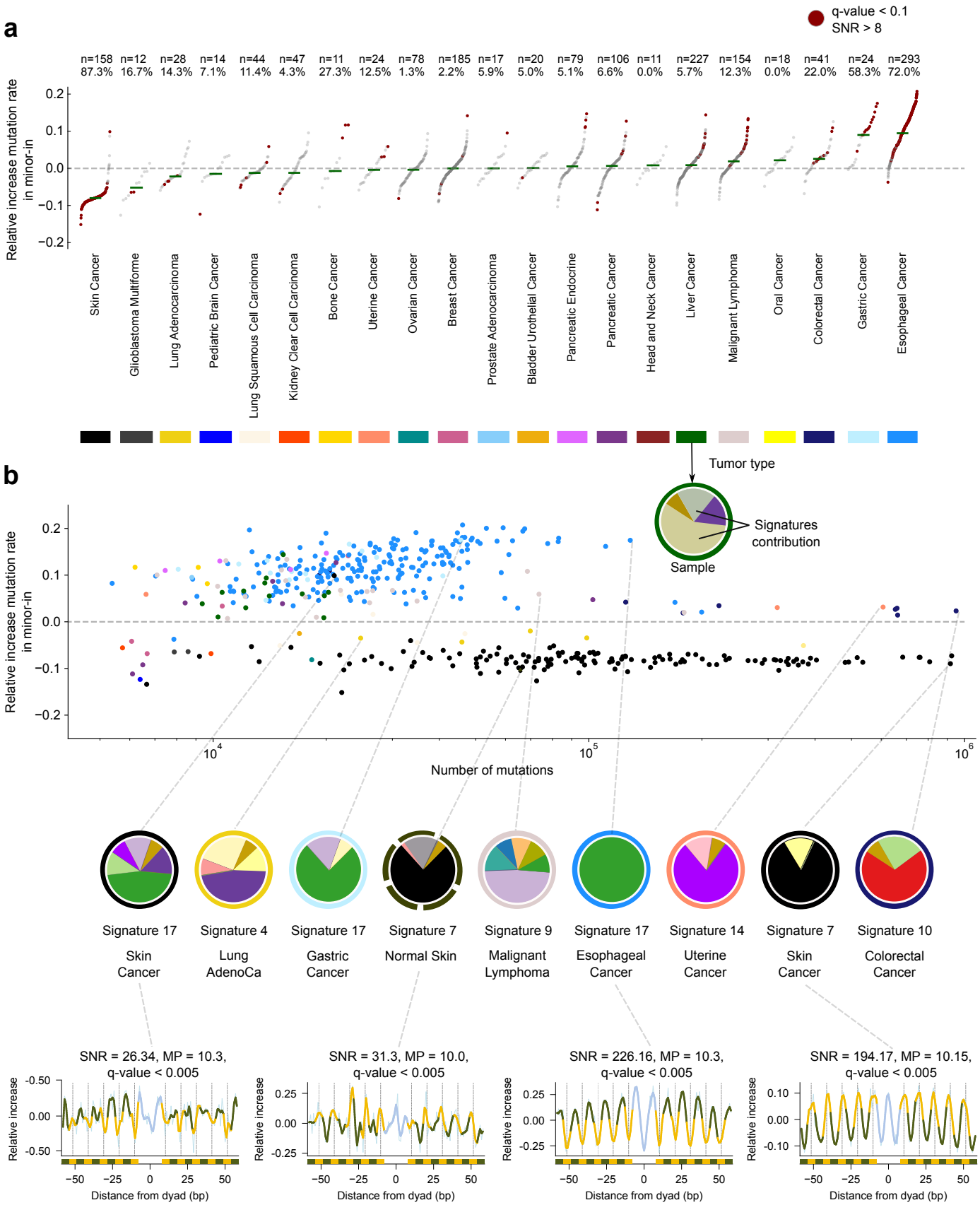

a
q-value < 0.1
SNR > 8
n=158 87.3%
n=12 16.7%
n=28 14.3%
n=14 7.1%
n=44 11.4%
n=47 4.3%
n=11 27.3%
n=24 12.5%
n=78 1.3%
n=185 2.2%
n=17 5.9%
n=20 5.0%
n=79 5.1%
n=106 6.6%
n=11 0.0%
n=227 5.7%
n=154 12.3%
n=18 0.0%
n=41 22.0%
n=24 58.3%
n=293 72.0%
Relative increase mutation rate in minor-in
0.2
0.1
0.0
−0.1
−0.2
Skin Cancer
Glioblastoma Multiforme
Lung Adenocarcinoma
Pediatric Brain Cancer
Lung Squamous Cell Carcinoma
Kidney Clear Cell Carcinoma
Bone Cancer
Uterine Cancer
Ovarian Cancer
Breast Cancer
Prostate Adenocarcinoma
Bladder Urothelial Cancer
Pancreatic Endocrine
Pancreatic Cancer
Head and Neck Cancer
Liver Cancer
Malignant Lymphoma
Oral Cancer
Colorectal Cancer
Gastric Cancer
Esophageal Cancer
b
Tumor type
Signatures contribution
Sample
Relative increase mutation rate in minor-in
Number of mutations
Signature 17
Skin Cancer
Signature 4
Lung AdenoCa
Signature 17
Gastric Cancer
Signature 7
Normal Skin
Signature 9
Malignant Lymphoma
Signature 17
Esophageal Cancer
Signature 14
Uterine Cancer
Signature 7
Skin Cancer
Signature 10
Colorectal Cancer
SNR = 26.34, MP = 10.3, q-value < 0.005
SNR = 31.3, MP = 10.0, q-value < 0.005
SNR = 226.16, MP = 10.3, q-value < 0.005
SNR = 194.17, MP = 10.15, q-value < 0.005
Relative increase
Distance from dyad (bp)

Figure 4

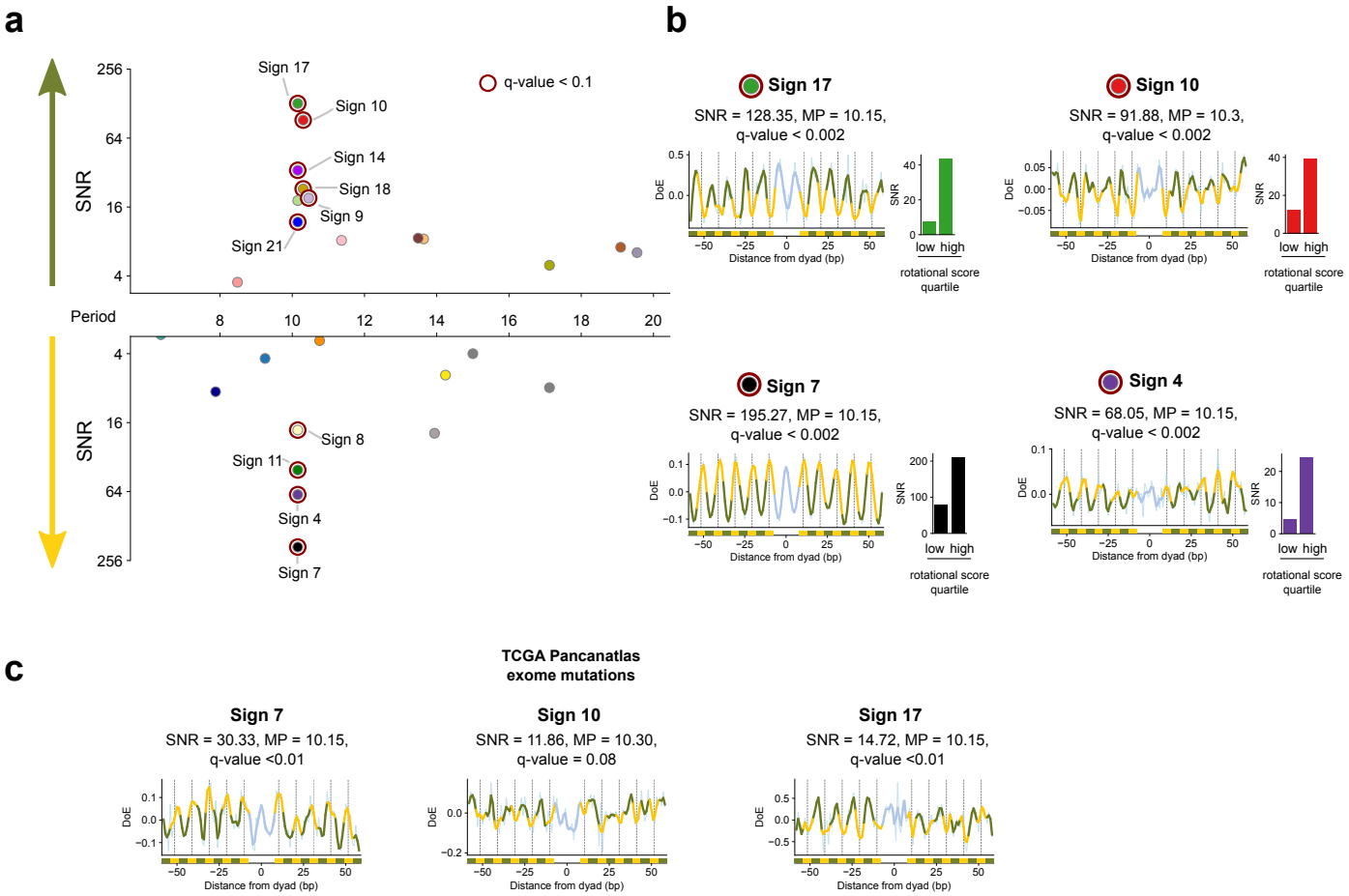

a
b
c
SNR
Period
q-value < 0.1
Sign 17
Sign 10
Sign 14
Sign 18
Sign 9
Sign 21
Sign 8
Sign 11
Sign 4
Sign 7
Sign 17
SNR = 128.35, MP = 10.15,
q-value < 0.002
Sign 10
SNR = 91.88, MP = 10.3,
q-value < 0.002
Sign 7
SNR = 195.27, MP = 10.15,
q-value < 0.002
Sign 4
SNR = 68.05, MP = 10.15,
q-value < 0.002
DoE
Distance from dyad (bp)
low high
rotational score quartile
TCGA Pancanatlas exome mutations
Sign 7
SNR = 30.33, MP = 10.15,
q-value <0.01
Sign 10
SNR = 11.86, MP = 10.30,
q-value = 0.08
Sign 17
SNR = 14.72, MP = 10.15,
q-value <0.01

Figure 5

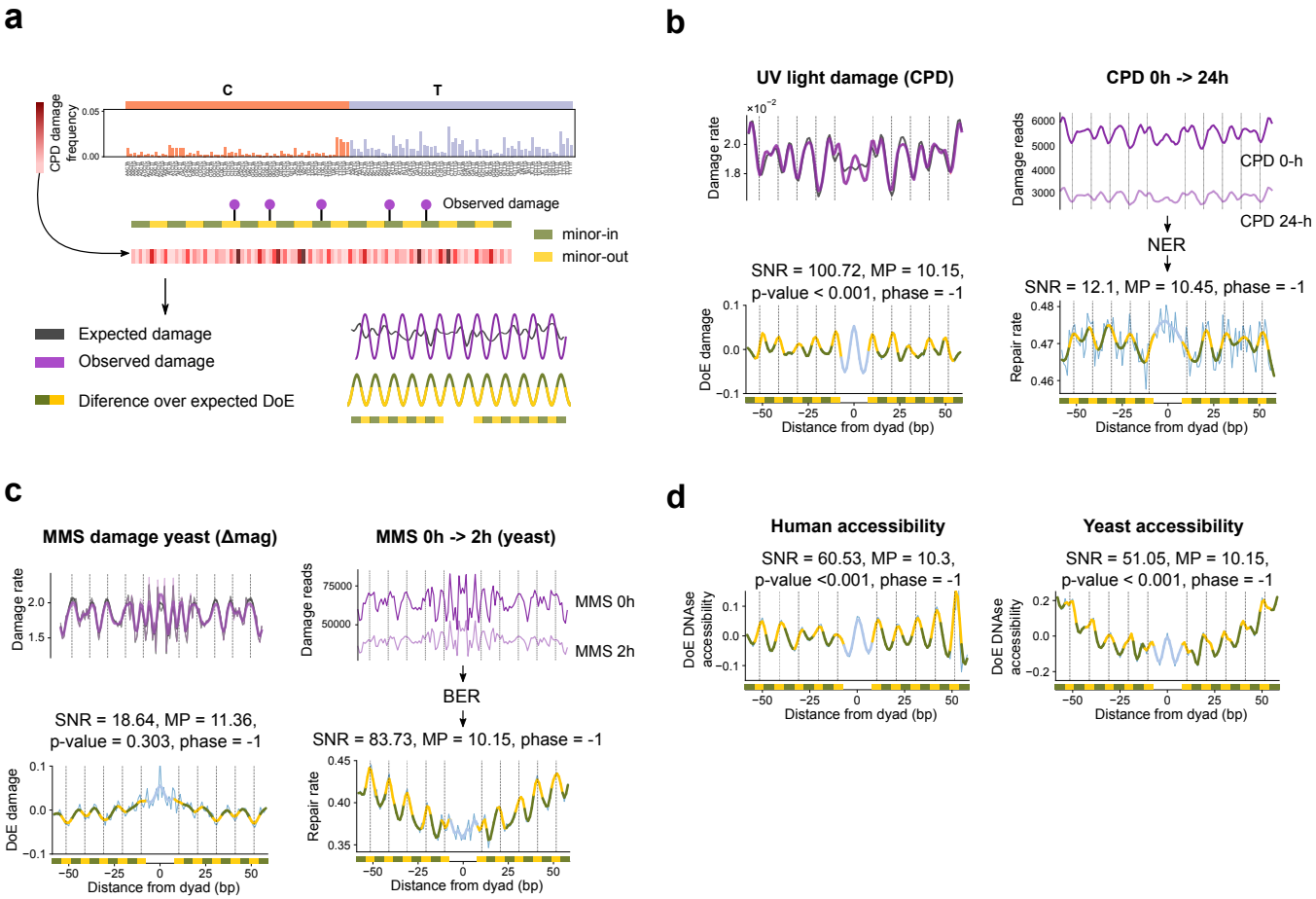

a
C
T
CPD damage frequency
0.05
0.00
Observed damage
minor-in
minor-out
Expected damage
Observed damage
Diference over expected DoE
b
UV light damage (CPD)
Damage rate
×10⁻²
2.0
1.8
SNR = 100.72, MP = 10.15,
p-value < 0.001, phase = -1
DoE damage
0.1
0.0
−0.1
−50
−25
0
25
50
Distance from dyad (bp)
CPD 0h -> 24h
Damage reads
6000
5000
4000
3000
CPD 0-h
CPD 24-h
NER
SNR = 12.1, MP = 10.45, phase = -1
Repair rate
0.48
0.47
0.46
Distance from dyad (bp)
c
MMS damage yeast (Δmag)
Damage rate
2.0
1.5
SNR = 18.64, MP = 11.36,
p-value = 0.303, phase = -1
DoE damage
0.1
0.0
−0.1
−50
−25
0
25
50
Distance from dyad (bp)
MMS 0h -> 2h (yeast)
Damage reads
75000
50000
MMS 0h
MMS 2h
BER
SNR = 83.73, MP = 10.15, phase = -1
Repair rate
0.45
0.40
0.35
−50
−25
0
25
50
Distance from dyad (bp)
d
Human accessibility
SNR = 60.53, MP = 10.3,
p-value <0.001, phase = -1
DoE DNAse accessibility
0.1
0.0
−0.1
−50
−25
0
25
50
Distance from dyad (bp)
Yeast accessibility
SNR = 51.05, MP = 10.15,
p-value < 0.001, phase = -1
DoE DNAse accessibility
0.2
0.0
−0.2
−50
−25
0
25
50
Distance from dyad (bp)

Figure 6

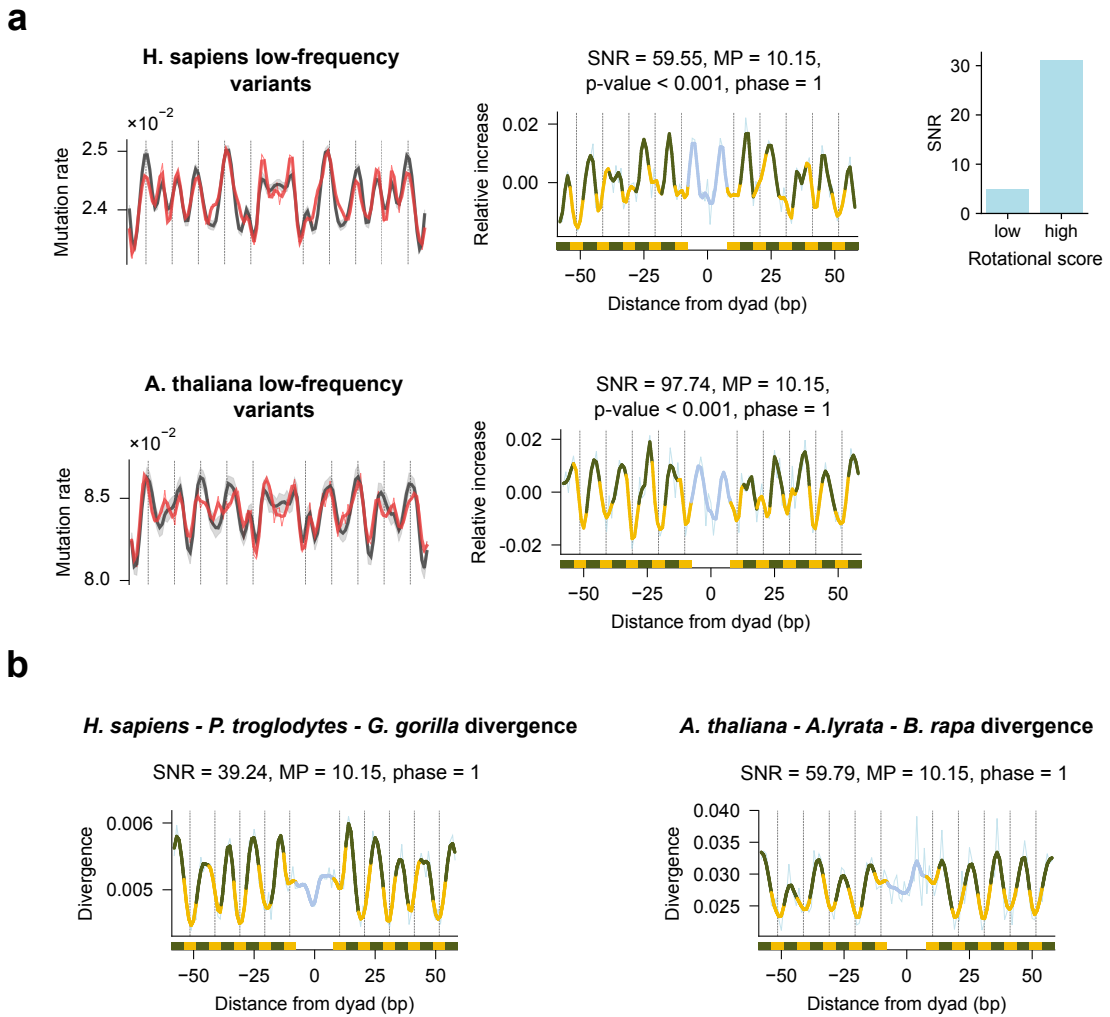
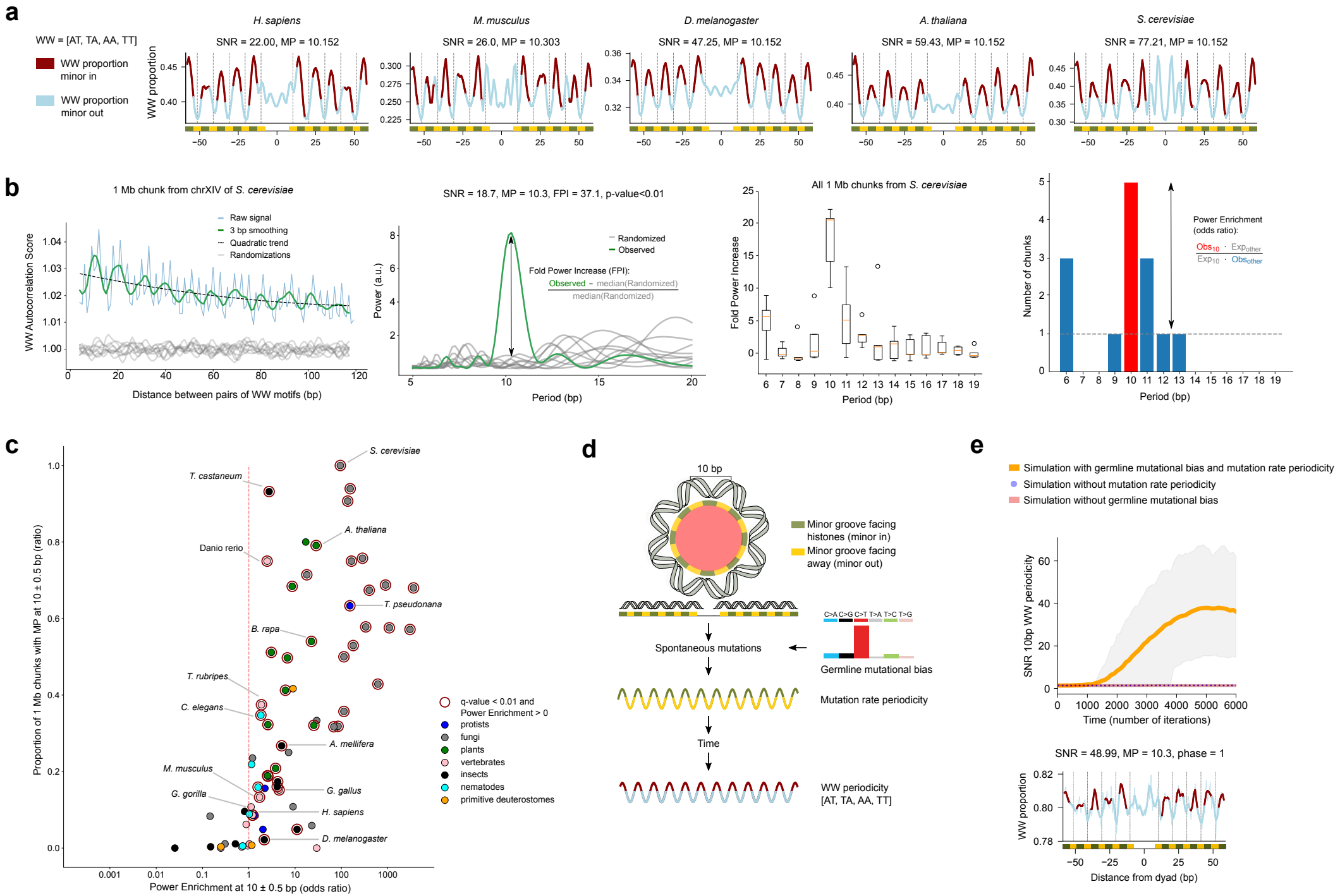

a
H. sapiens low-frequency variants
SNR = 59.55, MP = 10.15, p-value < 0.001, phase = 1
Mutation rate
Relative increase
Distance from dyad (bp)
SNR
Rotational score
low
high
A. thaliana low-frequency variants
SNR = 97.74, MP = 10.15, p-value < 0.001, phase = 1
b
H. sapiens - P. troglodytes - G. gorilla divergence
SNR = 39.24, MP = 10.15, phase = 1
A. thaliana - A.lyrata - B. rapa divergence
SNR = 59.79, MP = 10.15, phase = 1
Divergence

Figure 7


a
WW = [AT, TA, AA, TT]
WW proportion minor in
WW proportion minor out
WW proportion
H. sapiens
SNR = 22.00, MP = 10.152
M. musculus
SNR = 26.0, MP = 10.303
D. melanogaster
SNR = 47.25, MP = 10.152
A. thaliana
SNR = 59.43, MP = 10.152
S. cerevisiae
SNR = 77.21, MP = 10.152
b
1 Mb chunk from chrXIV of S. cerevisiae
Raw signal
3 bp smoothing
Quadratic trend
Randomizations
WW Autocorrelation Score
Distance between pairs of WW motifs (bp)
SNR = 18.7, MP = 10.3, FPI = 37.1, p-value<0.01
Randomized
Observed
Fold Power Increase (FPI):
Observed − median(Randomized)
median(Randomized)
Power (a.u.)
Period (bp)
All 1 Mb chunks from S. cerevisiae
Fold Power Increase
Period (bp)
Power Enrichment (odds ratio):
Number of chunks
Period (bp)
c
Proportion of 1 Mb chunks with MP at 10 ± 0.5 bp (ratio)
Power Enrichment at 10 ± 0.5 bp (odds ratio)
S. cerevisiae
T. castaneum
A. thaliana
Danio rerio
T. pseudonana
B. rapa
T. rubripes
C. elegans
A. mellifera
M. musculus
G. gallus
G. gorilla
H. sapiens
D. melanogaster
q-value < 0.01 and Power Enrichment > 0
protists
fungi
plants
vertebrates
insects
nematodes
primitive deuterostomes
d
10 bp
Minor groove facing histones (minor in)
Minor groove facing away (minor out)
C>A C>G C>T T>A T>C T>G
Spontaneous mutations
Germline mutational bias
Mutation rate periodicity
Time
WW periodicity [AT, TA, AA, TT]
e
Simulation with germline mutational bias and mutation rate periodicity
Simulation without mutation rate periodicity
Simulation without germline mutational bias
SNR 10bp WW periodicity
Time (number of iterations)
SNR = 48.99, MP = 10.3, phase = 1
WW proportion
Distance from dyad (bp)